

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

خوارزميات الذكاء الاصطناعي في تحليل النص العربي

مباحث لغوية ٦١

تحرير:

د. عبدالله بن يحيى الفيفي

تأليف:

د. وليد بن عبدالله الصانع

د. عبدالله بن صالح الراجح

د. فارس بن صالح القنيعير

د. عبدالرحمن بن محمد العصيمي

د. أيمن بن أحمد الغامدي

مباحث لغوية ٦١

خوارزميات الذكاء الاصطناعي في تحليل النص العربي

تأليف:

د. وليد بن عبدالله الصانع
د. عبدالله بن صالح الراجح
د. فارس بن صالح القنيعير
د. عبدالرحمن بن محمد العصيمي
د. أيمن بن أحمد الغامدي

تحرير:
د. عبدالله بن يحيى الفيفي

١٤٤١هـ - ٢٠١٩م





خوارزميات الذكاء الاصطناعي
في تحليل النص العربي

الطبعة الأولى
١٤٤١ هـ - ٢٠١٩ م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٤١هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
الفيفي، عبدالله بن يحيى
خوارزميات الذكاء الاصطناعي في تحليل النص العربي./ عبدالله بن يحيى الفيفي.- الرياض، ١٤٤٠هـ
..ص؛ .. سم
ردمك: ٠ -٦٠ -٨٢٢١ -٦٠٣ -٩٧٨
١ - اللغة العربية - معالجة البيانات أ. العنوان
ديوي ٤١,٢٨٥ ١١٣٣٥/ ١٤٤٠
رقم الإيداع: ١١٣٣٥/ ١٤٤٠
ردمك: ٠ -٦٠ -٨٢٢١ -٦٠٣ -٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
WOJOOOH
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com



هذا المشروع

مشروع تأليف سلسلة كتب في مجال (حوسبة العربية) يهدف إلى بناء تراكمٍ معرفي في مجال حيوي مهم، هو مجال (حوسبة العربية) . ويعد هذا الكتاب واحدا من سلسلة كتب صدرت في المركز.

يقع هذا المشروع ضمن سلسلة (مباحث لغوية) التي يشرف المركز على اختيار عنواناتها، وتكليف المحررين والمؤلفين، ومتابعة التأليف حتى إصدار الكتب. وهي سلسلة يجتهد المركز أن تكون سداداً لحاجات بحثية وعلمية تحتاج إلى تنبيه الباحثين عليها، أو تكثيف البحث فيها.

ويعدّ هذا الكتاب واحداً من كتب ثلاثة مترابطة في مشروع علمي واحد متخصص في (الذكاء الاصطناعي) :

١. العربية والذكاء الاصطناعي.
٢. تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية.
٣. خوارزميات الذكاء الاصطناعي في تحليل النص العربي.

مدير مشروع (العربية والذكاء الاصطناعي)

د.عبدالله بن يحيى الفيفي

كلمة المركز

يعمل المركز في مجال البحث العلمي ونشر الكتب مستهدفاً التركيز على المجالات البحثية التي ما زالت بحاجة إلى تسليط الضوء عليها، وتكثيف البحث فيها، ولفت أنظار الباحثين والجهات الأكاديمية إلى أهمية استثمارها بمختلف وجوه الاستثمار، وذلك مثل مجال (التخطيط اللغوي) و (العربية في العالم) و(الأدلة والمعلومات) و (تعليم العربية لأبنائها أو لغير الناطقين بها) إلى غير ذلك من المجالات، وإن من أهم مجالات البحث المستقبلية في اللغة العربية مجال (العربية والحوسبة ، والذكاء الاصطناعي) حيث إن اللغات الحية مرهونة حياتها مستقبلا بمدى تجاوبها مع التطورات التقنية والعالم الافتراضي، وكثافة المحتوى الالكتروني المكتوب، وهو ما يشكّل تحديا حقيقيا أمام اللغات غير المنتجة للمعرفة أو للتقنية.

وقد عمل المركز على تسليط الضوء على هذا المجال التخصصي؛ مستعينا بالكفاءات القادرة من المهتمين بالتخصص البيني (بين اللغة والحاسوب) مقدّرا جهودهم، وهادفاً إلى نشرها، وتعميم مبادئها، راغباً أن يكون هذا المسار العلمي مقررا في الجامعات في كلية العربية والحاسوب، ومجالا بحثيا يقصده الباحثون الأكديميون، والجهات البحثية العربية.

وقد أصدر المركز سابقا ستة عشر كتاباً مختصا في (حوسبة العربية) وفي الإفادة من (المدونات اللغوية) في الأبحاث العربية، ويحتفل بإصدار سبعة كتب جديدة مختصة في (حوسبة العربية والذكاء الاصطناعي)، ويقدمها للقارئ العربي، وللجهات الأكاديمية؛ للإفادة منها واعتماد ما تراه منها مناسباً لتعليمه والبناء عليه، وهذه الكتب السبعة هي: (تطبيقات الذكاء الاصطناعي في المعالجة الآلية، تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية، خوارزميات الذكاء الاصطناعي في تحليل النص العربي، مقدمة في حوسبة اللغة العربية، الموارد اللغوية الحاسوبية، المعالجة الآلية للنصوص العربية، تطبيقات أساسية في المعالجة الآلية للغة العربية).

ويشكر المركز السادة مؤلفي الكتب، ومحرريها، لما تفضلوا به من عمل علمي رصين، وأدعو الباحثين والمؤلفين إلى التواصل مع المركز لاستكمال المسيرة، وتفتيق فضاءات المعرفة.

وفق الله الجهود وسدد الرؤى.

الأمين العام

أ. د. محمود إسماعيل صالح

مقدمة المحرر[1]

الحمد لله رب العالمين، والصلاة والسلام على أشرف المرسلين، نبينا محمد وعلى آله وصحابته أجمعين، وبعد:

فأود أولاً أن أعبر عن وافر امتناني لمركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية على اهتمامه بإصدار سلسلة حول الذكاء الاصطناعي واللغة العربية، إذ شرفني بإدارة مشروع هذه السلسلة وتحرير أحد إصداراتها. وإنه لأمر يبعث على الغبطة والسرور أن نرى إصدارات عربية في مثل هذه الموضوعات التخصصية البينية التي تندر مراجعها في مكتبتنا العربية، خصوصاً تلك المراجع التي يتسم محتواها بالشرح المبسط لغير المتخصص مع ما تقدمه من ثراء وغنى في المعلومة، وهو السهل الممتنع

١- **عبدالله بن يحيى الفيفي**: أستاذ اللغويات الحاسوبية المساعد في جامعة الإمام محمد بن سعود الإسلامية في الرياض. درس البكالوريوس في اللغة العربية في جامعة الملك خالد في أبها، والماجستير في تعليم اللغة بمساعدة الحاسب في قسم اللغويات في جامعة Essex، والدكتوراه في اللغويات الحاسوبية في قسم الحاسب الآلي في جامعة Leeds، وكلاهما في بريطانيا. له العديد من الأبحاث المنشورة حول تقنيات معالجة اللغة العربية آلياً، والمدونات اللغوية وبرامجها الحاسوبية، وكذلك مدونات المتعلمين، وصناعة المعاجم الحاسوبية لمتعلمي اللغة العربية، إضافة إلى مشاركته في تأليف بعض الكتب المتخصصة في اللسانيات الحاسوبية، والمدونات اللغوية وتطبيقاتها. عمل محكماً لدى عدد من الدوريات العلمية والمؤتمرات الدولية. أنشأ المدونة اللغوية لمتعلمي اللغة العربية Arabic Learner Corpus، شارك في العديد من المشاريع العلمية والبحثية الوطنية في مجال تخصصه.

الذي نحتاجه في مثل هذه المؤلفات التي تثري مكتبتنا العربية بلا شك وتقدم المعرفة الحديثة في قالب يؤمل منه جذب أكبر عدد ممكن من المهتمين لهذه المجالات التخصصية الخصبة، التي باتت ميداناً للدراسة والبحث النظري إضافة إلى التجارب والتطبيقات العملية التي تتنافس عليها كبريات الشركات التقنية، وكذلك الجامعات والمراكز البحثية.

وتحسب لمركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية مبادرته في تبني مثل هذا المشروع وغيره من مشروعات السلاسل التي تعالج موضوعات متخصصة، وتفتح آفاقاً للقارئ العربي للحاق بركب العلم والمعرفة والاطلاع على آخر مستجداته.

وقد حرص المشاركون في تأليف هذا الكتاب - وهم نخبة من أساتذة الجامعات المتخصصين في ميدان الذكاء الاصطناعي ومعالجة اللغة العربية (مع حفظ الألقاب العلمية لهم) - على أن يكون الطرح تعليمياً متدرجاً مع شرح المصطلحات قدر الإمكان، وتقريب المعلومات للقارئ بأمثلة واضحة تساعد على الفهم والتطبيق، إلا أنه موجه بالدرجة الأولى لمن لديه مقدمة يسيرة عن تطبيقات الذكاء الاصطناعي Artificial Intelligence ومعالجة اللغة الطبيعية Natural Language Processing، وبناء الخوارزميات Algorithm؛ وذلك لتعميق معرفته حول خوارزميات الذكاء الاصطناعي التي يمكن الاستفادة منها في مجال تحليل النص العربي ومعالجة اللغة العربية التي تختلف في تركيبتها الصرفية والنحوية والدلالية عن اللغات اللاتينية التي حظيت باهتمام كبير في هذا الجانب، فهو -أي النص العربي- بحاجة إلى مزيد من البحث والدراسة لتكييف الخوارزميات المستعملة بما يتناسب مع خصائصه وقواعده، وهذا ما يحاول الكتاب شرحه باستعراض لعدة موضوعات حظيت بأبحاث عميقة في الآونة الحديثة. وفيما يلي عرض موجز لمحتويات الكتاب اعتماداً على الملخصات التي سترد لاحقاً في بداية كل فصل من فصوله.

ففي الفصل الأول يتحدث وليد الصانع عن طرق ومستويات معالجة اللغة في الذكاء الاصطناعي، مبيناً أن حوسبة معالجة اللغة تهدف إلى محاكاة الذكاء البشري؛ إذ إن اللغات البشرية تعتبر أحد أكثر الأنظمة تعقيداً، وهي تمر بمستويات عدة بدءاً من الصوت وانتهاءً بالخطاب. ويلقي هذا الفصل الضوء على مستويات معالجة

اللغة، مع استعراض بعض من الطرق المشهورة المستخدمة في معالجة اللغة في الذكاء الاصطناعي، ومنها على سبيل المثال خوارزميات تعلُّم الآلة (Machine Learning)، ونماذج ماركوف الخفية (Hidden Markov Models –HMMs)، والتعرُّف النمطي (Pattern Recognition) في الفضاء الدلالي، ونحوها مما يعطي القارئ لمحة عن طرق الذكاء الاصطناعي المستعملة في معالجة اللغة.

في الفصل الثاني يتناول فارس القنيعير خوارزميات التعلم العميق وتطبيقاته في معالجة اللغة، والتي تعد امتداداً لخوارزميات الشبكات العصبية. ويرجع سبب استخدام خوارزميات التعلّم العميق إلى قدرتها على تعلّم نماذج بالغة التعقيد كان من الصعب تعلمها سابقاً، وهذا أتاح العديد من التطبيقات التي تعالج احتياجات واقعية، منها معالجة اللغات الطبيعية. فيبدأ هذا الفصل بتقديم موجز عن الشبكات العصبية والتعلم العميق، ثم يتطرق لأهم المعماريات المستخدمة، وفي النهاية يعرض بعض تطبيقاتها في معالجة اللغات الطبيعية؛ للخروج بفهم عام عن خوارزميات التعلم العميق وكيفية تطبيقها في مجال معالجة اللغات.

وفي الفصل الثالث يتحدث عبدالله الراجح عن الترجمة الآلية، التي تعد من أصعب المشاكل في مجال الذكاء الاصطناعي، إذ تتطلب معارف لغوية متعددة لمحاكاة عمل المترجم المختص، ومع ذلك فهي تشهد تطوراً ملحوظاً في أداء أنظمتها بعد عقود من البحث والتطوير، وخصوصاً بعد تحولها من منهج الترجمة الآلية الإحصائية (Statistical Machine Translation) الذي كان مهيمناً على هذا الميدان لعدة عقود، إلى أن تحول المجتمع البحثي حديثاً وتبعته كبريات الشركات إلى المنهج المعتمد على الشبكات العصبية (Neural Machine Translation)، ويمكن اعتبارها نقطة التحول التي دخلت معها الترجمة الآلية عصراً جديداً، إذ يقدم الفصل الحالي عرضاً لأبرز ملامح هذا العصر، مع التطرق لبعض التحديات التي تواجه هذا المنهج البحثي الجديد.

في الفصل الرابع يتناول عبدالرحمن العصيمي نمذجة الكلمة العربية، إذ تمثل الكلمة ركيزة مهمة في فهم واستيعاب الخطاب المكتوب. ويهدف هذا الفصل إلى تزويد غير المتخصص بمقدمة لفهم أحدث الخوارزميات المستخدمة في بناء النماذج الحاسوبية للكلمة العربية الفصيحة المكتوبة. كما يحاول تفسير أسباب الصعوبات التي تكتنف نمذجة

الكلمة العربية تحديدًا، بدءًا بنظامها الصرفي الغير خطي ومرورًا بغناها الصرفي وانتهاءً بمستويات الغموض العالية في النص العربي. كما يقدم نمطين مشهورين لتحليل الكلمة: اللغوي والتوزيعي، ويقارن بينهما، وذلك عبر مقدمة لكل نمط وتحليل الخوارزميات المستخدمة وأشهر الأدوات المتاحة. وفي الختام، يسلط الضوء على أوجه القصور في بعض الخوارزميات عند تحليل ونمذجة اللغة العربية، والوسائل مقترحة لمعالجتها.

في الفصل الخامس يقدم أيمن الغامدي استعراضاً لتقنيات الذكاء الاصطناعي والمعالجة الحاسوبية للمتلازمات اللفظية والتراكيب الاصطلاحية، من خلال تتبع أهم الدراسات التي اهتمت بالمعالجة الحاسوبية لهذه الظاهرة اللغوية، إذ يبدأ الفصل بمقدمة تبين أهمية دراسة هذه الظاهرة وأهم مجالات البحث فيها، ثم يقدم إطاراً نظرياً مشتملاً على أهم الخصائص اللغوية المميزة لها في اللغة العربية، بالإضافة إلى استعراض أهم التصنيفات المستعملة للتراكيب الاصطلاحية في مستويات لغوية متعددة. بعد ذلك يستعرض أهم تطبيقات المعالجة الحاسوبية لهذه الظاهرة والتي تلخص المشاكل البحثية الرئيسة التي تتضمن التراكيب الاصطلاحية في أدبيات معالجة اللغات، كما يسلط الضوء بشكل خاص على مهمتي الاستخراج والتعرف الآلي، قبل أن يختم بعرض موجز لأبرز التحديات التي لا زالت تشكل عقبة في سبيل الوصول إلى درجات عالية من الدقة في مهام المعالجة الحاسوبية المختلفة لهذه الظاهرة اللغوية المعقدة.

ختاماً، أتقدم بالشكر الوافر - بعد شكر الله عز وجل - إلى مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية على ما قدمه للمحرر ولفريق التأليف من دعم متصل وتذليل للعقبات في سبيل تأليف هذا الكتاب الذي يؤمل أن يكون مرجعاً للمهتمين بهذا الميدان. كما أتقدم بالشكر الجزيل لجميع الزملاء المشاركين في تأليف فصول هذا الكتاب الذين بذلوا أوقاتهم وقدموا خلاصة أبحاثهم في مجالات تخصصهم، فلهم مني جزيل الشكر والامتنان.

المحرر / عبدالله بن يحيى الفيفي

الرياض- ٨ ذو القعدة ١٤٤٠هـ
ayjfaifi@gmail.com

موضوعات فصول الكتاب

الفصل الأول

طرق ومستويات معالجة اللغة في الذكاء الاصطناعي

د. وليد بن عبدالله الصانع

ملخص الفصل

تعتبر اللغات البشرية أحد أكثر الأنظمة تعقيداً والتي تميز الإنسان عن غيره من المخلوقات. وتمثل قدرة الإنسان على فهم وتوليد اللغة عاملاً من عوامل تميزه العقلاني. ولهذا، فإن حوسبة معالجة اللغة تعتبر أحد أهم تطبيقات الذكاء الاصطناعي والذي يهدف إلى محاكاة الذكاء البشري. وتمر معالجة اللغة في الذكاء الاصطناعي بمستويات عدة، بدءاً من الصوت وانتهاءً بالخطاب. ويعمل الباحثون على تطوير طرق متعددة لمعالجة اللغة في كل هذه المستويات. وفي هذا الفصل، نلقي الضوء على مستويات معالجة اللغة، وكذلك نستعرض بعضاً من الطرق المشهورة المستخدمة في معالجة اللغة في الذكاء الاصطناعي.

د. وليد بن عبدالله الصانع

أستاذ بحث مساعد في المركز الوطني لتقنية الذكاء الاصطناعي والبيانات الضخمة بمدينة الملك عبدالعزيز للعلوم والتقنية. حصل على درجتي البكالوريوس والماجستير في علوم الحاسب الآلي من جامعة الملك سعود بالرياض. قرأ الدكتوراه في مجموعة الذكاء الاصطناعي بجامعة يورك ببريطانيا. عمل سابقا مهندساً للبرمجيات في قسم الأبحاث والتطوير في شركة الإلكترونيات المتقدمة، ثم مهندساً للنظم والبرمجيات في شركة الاتصالات السعودية قبل أن ينتقل للعمل باحثاً في مدينة الملك عبدالعزيز للعلوم والتقنية. تتمثل اهتماماته البحثية في تعلم الآلة، وتحديداً في تعلم البرامج المنطقية، تعلم النماذج الاحتمالية، وتعلم البرامج المنطقية الاحتمالية، وتطبيقات تعلم الآلة في نمذجة ومعالجة اللغة.

١ – المقدمة

اللغة هي الوسيلة التي يُعبِّر بها الإنسان عمَّا يدور في عقله من تصورات وأفكار. وهي نظام ترميزي مُتعارَف عليه بين مجموعة من البشر يستخدمونه لإيصال الأفكار والصور التي في عقولهم بحيث يستطيع المستقبل[1] لهذه الرموز من نفس المجموعة أن يحولها في عقله لنفس الأفكار والصور التي أراد المتكلم إيصالها، أو قريبًا منها. إذ إن اللغة البشرية هي تمثيل لأفكار ومفاهيم بطريقة مسموعة (الكلام) أو مكتوبة (النص). وبناءً عليه، فإن استخدام الإنسان للغة مبني على ثلاث قدرات أساسية وهبها الله سبحانه وتعالى له، وهي:

- **قدرته على تعلم اللغة** (Language Acquisition): أي قدرته على استقراء (Induce) القواعد التركيبية والدلالية للوحدات والتراكيب اللغوية، كالقواعد الصرفية والنحوية ودلالات المفردات، من خلال الأمثلة التي يتعرض لها سماعًا (في بدايته كطفل)، أو قراءةً (بعد تعلمه القراءة) (Clark, 2002).

مثال:

يستمِعُ طفلٌ إلى الناس من حوله يقولون في حديثهم عن ذُكورٍ:

أعطيتُـ**ه**، حدثتُـ**ه**، سلمتُـ**ه**، أكرمتُـ**ه**، ...

وفي حديثهم عن إناثٍ:

أعطيتُـ**ها**، حدثتُـ**ها**، سلمتُـ**ها**، أكرمتُـ**ها**، ...

ويعلم هذا الطفل أن الكلمات تشير إلى أحداث مرتبطة بالزمن (أفعال) قام بها المتحدث تجاه أطراف ثالثة، ذكورًا وإناثًا. فعندئذ، يقوم باستقراء النظرية اللغوية التالية:

- إذا أراد متحدث الإشارة إلى فعل تجاه طرف ثالث ذكر فإنه يلحق «**ه**» بالفعل، وإذا أراد الإشارة إلى فعل تجاه طرف ثالث أنثى فإنه يلحق بالفعل «**ها**».

---

١ – سنستخدم كلمة «المُستقبِل» في مناسبات مختلفة خلال هذا الفصل للإشارة للسامع أو القارئ عندما لا يُحدد السياق هل المقصود كلام أم نص.

- **قدرته على استقبال اللغة، أو ما يعرف باللغة الاستقبالية** (Receptive Language): وهي القدرة على معالجة وفهم اللغة وفق القواعد اللغوية التي تعلمها ووَفْق المعتقدات (Beliefs) التي كوَّنها عن العالم (خصائص الموجودات وعَلاقاتها مع بعضها). أي تحويل الرموز والتراكيب اللغوية إلى المفاهيم العقلية المناسبة. وتستخدم هذه القدرة في معالجة ما يسمعه الإنسان أو يقرأه.

مثال:

استقبل شخصٌ هذه الجملة «أعطيتُ الرَّجُلَ الجالِسَ هناك كأسًا من الماء». تتمثَّل اللغة الاستقبالية بقدرة المُستقبِل على معالجة هذه الجملة، فأحد الفرضيات هي أن يقوم المُستقبِل بتفكيك الجملة إلى الكلمات المُكوِّنة لها كالتالي:

أعطى، ت، ال، رجل، ال، جالس، هناك، كأسا، من، ال، ماء.

وتحديد أدوارها في تركيب الجملة كالتالي:

أعطى (فعل ماض)، ت (حرف ينوب عن اسم)، الـ (كلمة تعريفية لما بعدها)، رجل (اسم جنس)، الـ (كلمة تعريفية لما بعدها)، جالس (صفة)، هناك (اسم إشارة)، كأسًا (اسم جنس)، من (حرف)، الـ (كلمة تعريفية لما بعدها)، ماء (اسم جنس).

يمكن بعد ذلك تحليل تركيب الجملة وَفْق قواعد التركيب اللغوية، بحيث تُحدد أولوية ترابط الكلمات مع بعضها البعض لتكوين العبارات انتهاءً بتكوين الجملة (Parsing). فعلى سبيل المثال، تُربط الكلمتان «الجالس» و «هناك» لتكوين العبارة «الجالس هناك» ومن ثم تُدخل عليهما كلمة «الرجل» لتكوين العبارة الأوسع «الرجل الجالس هناك»، وذلك لتحديد قراءة أن «الجالس هناك» عبارة مرتبطة بـ «الرجل». ويقوم المُستقبِل بدمج التحليل الذي توصل له مع دلالة المفردات (Lexical Semantics) ومع مُعتقداته، وهي حقيقة أن الكأس يُعطَى والرجل هو الذي يُعطِي، لاستخراج دلالة الجملة وتحويلها إلى المفاهيم العقلية المناسبة، وهي تحديد المُعطِي، والمُعطَى، والمُعطَى له، وصفة المُعطَى له أثناء الكلام.

- **قدرته على إنتاج اللغة، أو ما يعرف باللغة الإنتاجية** (Productive Language) **أو اللغة التعبيرية** (Expressive Language): وذلك بتحويل المفاهيم والتصوُّرات العقلية إلى تراكيب لغوية مناسبة تُوصِل هذه التصورات والمفاهيم، أو قريبًا منها، إلى المستقبل. وهي عملية عكسية للغة الاستقبالية. وتستخدم هذه القدرة أثناء الكلام أو الكتابة.

مثال:

يُريد المُتحدِّث إيصال مفهوم في عقله يتمثَّل في حادثةٍ انتهت، وهي إعطاء رجل يجلس الآن في مكانٍ يمكن رؤيته كأسًا من الماء. فأحد الفَرَضيَّات أنه يستدعي قواعد اللغة التي تعلَّمها لبناء الجُمل، ومن ثمَّ يستدعي الكلمات التي تُوصِل المعنى ويُولِّد الجملة. يُمكِن أن تتمَّ هذه العملية كالتالي:

- المفهوم المُراد إيصالُهُ يُشير إلى فِعل، ويوجد فيه فاعل ومفعولان، ووصف لحال أحد المفعولين، فيستدعي قاعدة لغوية تعلَّمها للتعبير عن هذا المفهوم ليحصل على:

  فعل + فاعل + المفعول الأول + صفة + المفعول الثاني.[1]

- يبحث عن كلمةٍ في الذاكرة تُوصِل معنى الحدث، وهي هنا الإعطاء، ثم يصرفها لتُناسب الزمان الماضي ولتشير إلى أن مَن أعطى هو المُتحدِّث، ووَفْقًا لقواعد الصرف التي تعلَّمها، يولد الكلمة «أعطيت».
- يبحث في الذاكرة عن كلمة تُشير إلى مَن حدث له الفعل، فيولد كلمة «الرجل».
- يستدعي الكلمات المناسبة لوصف المُعطى له، ويُولد عبارة «الجالس هناك»[2].
- يُشار إلى المُعطى؛ وذلك باستدعاء الكلمات الدلالية من الذاكرة، وتوليد عبارة «كأسا من الماء»[3].

١– يشير الرمز «+» هنا إلى علاقة ترتيب بين الكلمات.

٢– هذه العبارة أيضا تولد وفقًا لقواعد تركيبية بنفس طريقة توليد الجملة، لذا فإننا لا نحتاج لإعادة شرحها مرة أخرى.

٣– نفس الحال الذي ذكر في توليد عبارة –الجالس هناك– ينطبق هنا أ أيضًا.

ولأن الذكاء الاصطناعي هو فنٌّ يهتم بدراسة وفهم الإدراك البشري، ومن ثَمَّ محاولة بناء برمجيات حاسوبية تُحاكي عملية الإدراك، فإن الباحثين في مجال الذكاء الاصطناعي يعكفون على دراسة هذه القُدُرات الثلاث لدى الإنسان ومحاولة بناء برمجيات حاسوبية تُحاكيها. وتبقى كيفية عمل هذه القُدُرات لدى الإنسان سِرًّا من أسرار الكون التي وضعها الله سبحانه وتعالى ولا سبيل لمعرفتها على سبيل اليقين[1]. وتستمد النظريات التي تطرح في كثير من الأدبيات الخاصة بهذه الدراسات من فروع مختلفة تُمثل بنية تحتية لمجال الذكاء الاصطناعي، ومن هذه الفروع: اللسانيات (Linguistics) واللسانيات النفسية (Psycholinguistics)، الرياضيات والإحصاء (Mathematics and Statistics)، الفلسفة والمنطق (Philosophy and Logic)، علم الإدراك (Cognitive Science)، نظرية الحوسبة (Theory of Computation). لذا فإن الدارس لمجال اللسانيات الحاسوبية يعمل في منطقة تقاطع لهذه الفنون، إضافة إلى فنون أخرى تمسُّ بعض التطبيقات، مثل معالجة الإشارات (Signal Processing) لِمَنْ يعمل على تحويل الكلام المسموع إلى نصوص مكتوبة.

وفي هذا الفصل سنتطرَّق إلى مستويات معالجة اللغة البشرية بدءًا من تكوين الكلمة من الأصوات إلى إدراك المعنى وبناء المعتقدات. ثم سنستعرض بعض الأمثلة على المواضيع التي يعمل عليها الباحثون في مجال اللسانيات الحاسوبية، مع التركيز على معالجة النص فقط دون معالجة الكلام. وآخِرًا سنستعرض بعضًا من الطرق المُستخدَمة لمعالجة النصوص.

٢- مستويات معالجة اللغة

تمرُّ معالجة اللغات الطبيعية على مستويات عدَّة، بدايةً من تكوين الكلمة من الأصوات، مرورًا بتكوين الجُمل من الكلمات، وانتهاءً بفهم الكلام. ولكل مُستوًى قواعدُ تركيبية تُستخدم لتكوين وحدات هذا المستوى. يمكن تحديد المستويات معالجة اللغة بالتالي (Allen J. , 1994; McCarthy, 2018):

١- هذا اعتقاد الكاتب على الأقل.

- **المستوى الصوتي** (Phonetic Level): وهو المستوى الأساسي (primitive) المُكوِّن للغة. وفي هذا المستوى، تُحلل الأصوات وترابطها مع بعضها لمعرفة الكلمات المُرادة.
- **المستوى الصرفي** (Morphological Level): في هذا المستوى تُحلل بنية الكلمات بناءً على وحدات أساسية، تُسمَّى الوحدات الصرفية (Morphemes). فمثلا، كلمة «يذهبون» مُكوَّنة من ثلاث وحداتٍ، الأولى «يـ» للإشارة إلى أن الفعل قام به طرف ثالث، والثانية «ذهب» وهو الفعل، ويُمثل الوحدة الأساسية للكلمة، والثالثة «ون» للإشارة إلى جمع المُذكَّر.
- **المستوى التركيبي للجمل** (Syntactic Level): في هذا المستوى، يُحلل ترابط الكلمات لمعرفة كيف تتكون الجملة، ومن خلال هذا التحليل يُمكِن تحديد قراءة الجملة. فعلى سبيل المثال، يمكن تحليل جملة «رأيتُ الرَّجُلَ جالِسًا» على قراءتين، الأولى وهي الشاذة:

{[(رأى ت) (الـ رجل)] (جالسا) }

وفي هذه القراءة، تكون كلمة «جالسًا» حالا للرائي. لأنه وبحسب التحليل أعلاه، رُبطت الكلمتان «رأيت» و «الرجل» أولًا لتكوين عبارة «رأيت الرجل» ومن ثمَّ أدخلت كلمة «جالسًا» إلى هذه العبارة كما هو موضح في الأقواس. فيكون ناتج التحليل أن الذي رأى كان جالسًا وهو يرى الرجل. أما القراءة الثانية وهي الشائعة:

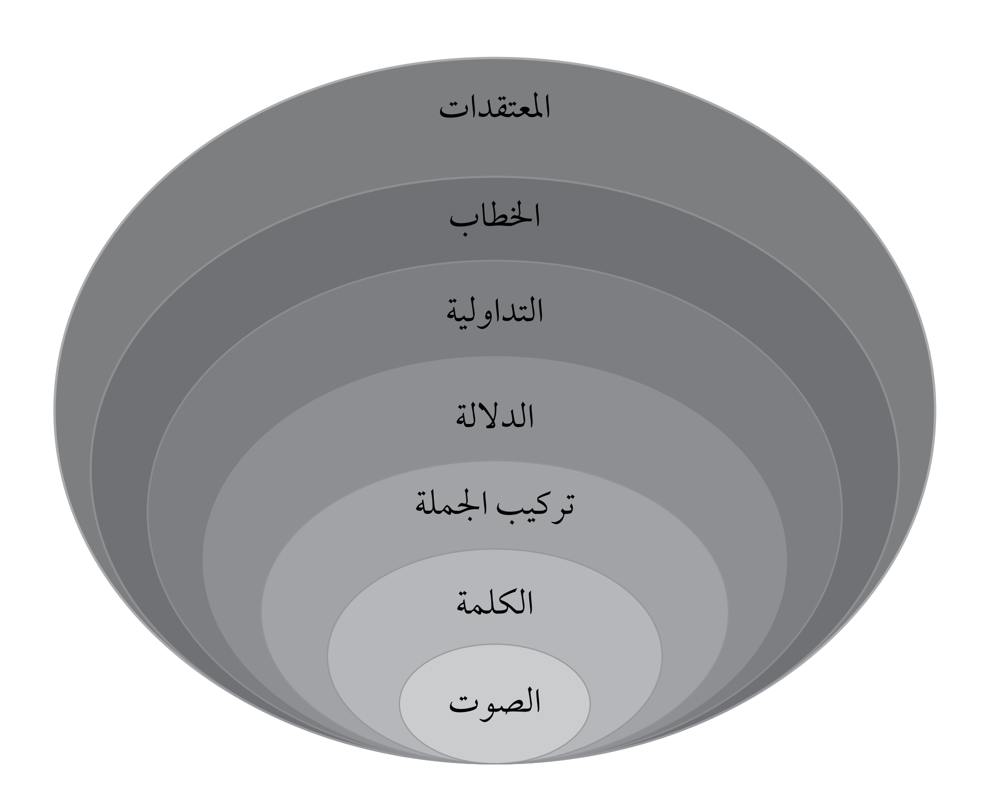
{ (رأى ت) [(الـ رجل) (جالسا)] }

ففيها رُبطت الكلمتان «الرجل» و «جالسًا» مع بعضهما أولا لتكوين عبارة «الرجل جالسًا»، ثم أُدخلت على هذه العبارة كلمة «رأيت» كما هو مُوضَّح في الأقواس. فتُشير القراءة إلى أن المرئيَّ هو الذي كان جالسًا، والمُتحدِّث رآه على هذه الحال.

وهذا الاختلاف في تحليل نفس الجملة يُعرَف بالغموض التركيبي (Syntactic Ambiguity) (Manning & Schütze, 1999)، أي أن الجملة يمكن تركيبها بأكثَرَ من طريقةٍ لكُلٍّ منها ناتجٌ مُختلِف.

- **المستوى الدلالي** (Semantic Level): في هذا المستوى، تُستخرج المعاني الحرفية للمفردات، ومن ثم تُربط هذه المعاني لتكوين معنى الجملة كاملة، وهو ما يعرف بالدلالة التركيبية (Compositional Semantics) (Sternefeld & Sternefeld, 2013). لا يُؤخذ سياق الجملة أو مناسبتها في الحُسبان لاستخراج معناها، وإنما يؤخذ المعنى الحرفي فقط. فجملة «بلغ السيل الزبى» تعني أن هنالك ماءً ارتفع ووصل الزبى.
- **المستوى التداولي** (Pragmatic Level): في هذا المستوى، تُحلل الجملة وَفْق السياق والمقام الزماني والمكاني والاجتماعي الذي قيلت فيه وليس بالضرورة أن يكون مطابقًا للمعنى الحرفي للكلمات المُكوِّنة لها. ولمعرفة معنى الجملة في هذا المستوى، يحتاج المُستقبِل إلى استخدام معرفة إضافية خارج إطار المعرفة اللغوية (Extra-linguistic Knowledge)، وهي المُعتقدات (Beliefs) التي لدى الشخص عن العالم. والتحليل على هذا المستوى ضروريٌّ لمعرفة المجاز اللغوي، والتعريض، والتلميح، وفهم الحِكَم والأمثال والقصائد. وفي هذا المستوى، على سبيل المثال، يُعرف المعنى المراد لجملة «بلغ السيل الزبى»، وهي أن الصبر قد نَفِدَ، إذا لا ماء ارتفع ولا زبى موجودة في الواقع المُشار إليه.
- **المستوى الخطابي** (Discourse Level): في هذا المستوى، يُحلل الخطاب بمعالجة العلاقات بين الجمل المكونة له. فتُحلل كل جملة وفق الجمل التي تسبقها لمعرفة تأثير تلك الجمل على وجود هذه الجملة. ويعمل تحليل المستوى الخطابي على معرفة الهدف العام من الكلام والرسائل التي يحتويها.

وكما هو مُلاحظ، فإن اللغة تُمثِّل نظامًا مُتعدِّد المستويات، كلُّ مُستوى يحتوي عناصر تتركَّب مع بعضها البعض لتكوين مركبات تمثل بدورها عناصر المستوى الذي فوقه كما هو موضح في الشكل ١ أدناه. كما أن المعالجة في كُلِّ مستوى قد تستخدم معلومات من المستوى الذي فوقه. فالإنسان يستخدم المعتقدات التي لديه لمعالجة اللغة في جميع المستويات بدءًا من الصوت وانتهاءً بالخطاب. وبعد المعالجة يحصل الإنسان على معتقدات إضافية تضاف لمعتقداته السابقة أو تقوم بتغييرها وهو ما يُعرَف بـتحديث المعتقدات (Belief Updating) أو مراجعة النظريات (Theory Revision).



الشكل ١: مستويات معالجة اللغة

١,٢ بعض عمليات مُعالجة اللغة في مختلف المستويات

يعمل الباحثون في مجال معالجة اللغات الطبيعية على تطوير أنظمةٍ لتحليل التراكيب اللغوية في مختلف المستويات المُوضَّحة في الشكل ١. إذ إنه من الضروري العملُ على المستويات الدُّنيا، مستويات الصوت والكلمة والجملة، لكي يمكن بناء أنظمةٍ تعالج اللغة في المستويات العليا، كمستويات التداولية والخطاب. وسنسلط الضوء في هذا الجزء على بعض من المهامِّ التي يعمل عليها الباحثون ومُطوِّرو أنظمة معالجة اللغات الطبيعية. وأود التنويه إلى أنني هنا لا أُغطي إلا جزءًا يسيرًا من هذه المهامِّ وذلك لإيضاح المفاهيم والأفكار الرئيسية فقط؛ إذ لا يمكن شرح تفاصيل هذه المهام في فصلٍ، ولا حتى في كتابٍ واحدٍ.

فعلى المستوى الصوتي، يعمل الباحثون على دراسة تحويل الموجات الصوتية إلى كلماتٍ مكتوبةٍ، وهو ما يعرف بتطبيقات التعرُّف على الكلام (Speech Recognition)، أو تحويل الكلام إلى نصوص (Speech-to-Text) (Gales & Young, 2007). ويتطلب العمل على هذا المستوى دراسة لنمطية الأصوات بحيث يمكن التعرُّف على الصوت بناءً على الأصوات المصاحبة له (معرفة صوتٍ بناءً على الأصوات التي قبله مثلًا). ومما يُساعد على التعرُّف على الكلام أيضًا دراسة نمطية تسلسل الكلمات؛ إذ إن دراسة هذه النمطية تساعد في معرفة الكلمة التي قيلت بناءً على الكلمات المصاحبة لها. ومن المهامِّ التي يعمل عليها الباحثون أيضًا توليد الكلام (Speech Synthesis) أو تحويل النصوص إلى كلام (Text-to-Speech) (Allen, Hunnicutt, Klatt, Armstrong, & Pisoni, 1987). ويهدف هذا العمل على تطوير أنظمةٍ تقرأ الملفات النصية.

أما على مستوى مُعالجة الكلمة، فإنَّ من أهم عمليات المعالجة في هذا المستوى هو التحليل الصرفي (Morphological Analysis) (Jurafsky & Martin, 2008; McCarthy, 2018). فمن خلال هذا التحليل، تُحدد الزوائد (Affixes) التي تدخل على الكلمة والتي يُمكِن أن تكون في بدايتها (Prefixes)، مثل حرف الياء في كلمة «يـذهب»، أو في وسطها (Infixes)، مثل حرف الألف في كلمة «ذاهب»، أو في نهايتها (Suffixes)، مثل حرفي الواو والألف في كلمة «ذهبوا». فمن خلال التحليل الصرفي للكلمات السابقة يمكن معرفة أن هذه الكلمات لها نفس الجذر (Root) وهو «ذهب». وتختلف المعالجة الصرفية باختلاف الهدف منها. ففي تطبيقات استرجاع المعلومات (Information Retrieval) مثلًا، وهو المُصطلح العلمي المستخدم للإشارة للتقنية التي تُبنى عليها مُحركات البحث (Search Engines)، يقوم بعض المُطوِّرين لهذه التطبيقات باستخراج جُذور الكلمات لنصوص الاستعلام (Queries) وذلك للبحث عن النصوص التي تحتوي على كلماتٍ لها نفس الجذور بدلًا من مطابقة الكلمات كما هي. فلو أدخل المستخدم مثلًا الاستعلام «أعمال الحجاج»، فإنه بمطابقة النصوص التي فيها إحدى هاتين الكلمتين كما هما، سيسترجع النظام تلك النصوص التي تحتوي إحدى هاتين الكلمتين أو كلاهما فقط، وسيستبعد نصوصاً لا تحتوي أيا منهما ولكن تحتوي على كلمات مثل «حج»، «حجيج»، «يحجُّون»، وغيرها من مشتقات «حجَّ». أما

بإعادة الكلمات في نص الاستعلام، وفي النصوص التي في قاعدة البيانات إلى جذورها، فإن أيَّ نص توجد فيه كلمة مُشتقَّة من «حجَّ» سيتمُّ استرجاعه وعَرْضه على المُستخدِم.

أما على المستوى التركيبي، فمن مهام المعالجة الأساسية التعرُّف على أجزاء الكلام (Parts of Speech) للكلمات (Kübler & Mohamed, 2011; Manning & Schütze, 1999). ومصطلح أجزاء الكلام غير مُستخدَم في دراسات اللغة العربية بشكلٍ شائع[1]، لكنه من المصطلحات المستخدمة في الدراسات المتعلقة ببعض اللغات الأخرى، وخاصة الإنجليزية. ويُشير مصطلح أجزاء الكلام إلى الأصناف التي يمكن أن تُنسَب إليها كلمات اللغة بناءً على دورها التركيبي. على سبيل المثال، يمكن اعتبار هذه القائمة: اسم جنس (ومثال ذلك كلمة «إنسان»)، اسم شخص (ومثال ذلك كلمة «محمد»)، فعل، ضمير، حرف، صفة، حال، رابط (مثال ذلك واو العطف)، أجزاء للكلام. ولا يوجد اتفاق تامٌّ على مجموعة ثابتة لأجزاء الكلام للغة ما، بل إن هذه المجموعة قد تتغيَّر بحسب نوع التحليل ورؤية من يقوم بذلك. فقد يبدأ بعض الدارسين بمجموعةٍ مُعيَّنة، ومِن ثَمَّ يقومون بإضافة أجزاء أخرى عند الحاجة. إحدى الجهات التي قامت بتبني قائمةٍ لأجزاء الكلام هو اتحاد البيانات اللغوية (Linguistic Data Consortium)[2] بجامعة بنسلفانيا بالولايات المتحدة الأمريكية. وتُستخدم هذه القائمة لأكثر من لغة ومن ضمنها اللغة العربية. وهناك باحثون آخرون يتبنُّون مجموعة مختلفة من أجزاء الكلام للغة العربية بحسب المهمة التي يعملون عليها، وآلية التحليل التي يستخدمونها.

أما المعالجة على المستوى الدلالي، فتتمثل في تحليل دلالة المفردات من خلال معرفة المفردات التي لها نفس المعنى، أو تلك التي لها معانٍ متقاربة (Landauer & Dumais, 1997; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). يمكن استخدام التحليل الدلالي في تطبيقات استرجاع المعلومات لاسترجاع النصوص التي تحتوي على كلماتٍ مُترادفةٍ للكلمات التي أدخلها المُستخدِم، أو لها معان قريبة منها.

١- بحسب علم الكاتب، أنه في اللغة العربية يُستخدم مصطلح «أقسام الكلام» للإشارة إلى الأقسام الرئيسية فقط، وهي الاسم والفعل والحرف.

2- https://www.ldc.upenn.edu/

فبالعودة إلى المثال السابق وهو الاستعلام باستخدام العبارة «أعمال الحُجَّاج»، يمكن من خلال التحليل الدلالي إعادة النصوص التي لا تتعلق بالحج فقط، بل حتى بتلك التي تتعلق بكلماتٍ قريبة منها دلاليا كـ «العُمرة» و «الطواف» و «السَّعْي»، وربما أيضًا تلك النصوص المتعلقة ببعض المشاعر كـ «مُزدلفة» و «مِنًى» و «عرفات»، بحكم قرب هذه المفردات دلاليًّا من الحج. تجدُرُ الإشارة إلى أن باحثين قاموا بتطوير قاعدة بيانات تحوي مفردات بعض اللغات، ومنها العربية، وارتباطاتها الدلالية من حيث الترادُف وقُرْب المعاني، والمفاهيم (Concepts) التي تحملها هذه الكلمات، والعلاقات بين هذه المفاهيم كالعَلاقة الهرمية، مثل العامِّ والخاصِّ («رجل:إنسان»)، وعلاقة الجزء من الكل («يد:جسم»). وتُعرَف هذه الشبكة بشبكة الكلمات (WordNet)[1] [2].

أما التحليل على المستوى التداولي، فهو من أكثر المهامِّ تحديًا. فبحسب معرفتي، أن الأبحاث في المعالجة على هذا المستوى محدودةٌ مقارنةً بالأبحاث التي تُعالج اللغة في المستويات الأخرى. ولعل التطبيقات التي تهدف لمعرفة مقاصد الجمل وما يبنى عليها (Textual Entailment) تعمل على هذا المستوى إضافة إلى المستوى الدلالي (Androutsopoulos & Malakasiotis, 2010).

وفيما يتعلق بتحليل الخطاب، فإنه يُستخدم لبناءِ كثيرٍ من التطبيقات، منها على سبيل المثال، التلخيص الآلي (Text Summarization) (Marcu, 2000)، وذلك بتحليل النص لاستخراج الجمل الأكثر أهميةً، والتي تُوصل المعنى الذي أراد الكاتب إيصاله. كذلك معرفة المُؤلِّف (Author Identification)، وذلك من خلال تحليل النصوص التي كتبها سابقًا لمعرفة أسلوب كتابته (Writing Style) ومقارنة هذا الأسلوب بالنص الذي تتم معالجته لمعرفة ما إذا كان هو كاتب هذا النص أم لا. ومن التطبيقات أيضا تحليل النصوص لمعرفة حقبها التاريخية وأحوال مؤلفيها عند كتابتها، وغيرها من التحليلات التي يقوم بها النُّقَّاد الأدبيون يدويًّا.

1- https://wordnet.princeton.edu/

2- http://globalwordnet.org/resources/arabic-wordnet/

٢,٢ طرق معالجة اللغة

تتمثل عملية المعالجة في كونها دالة[١] f تأخذ مجموعة من المُدخَلات[٢] $\boldsymbol{X}=\{x_i\}_{i=1}^{n}$ لتقوم ببعض العمليات عليها وتعيد قيمة $\boldsymbol{y}$ يمكن استخدامها لاتخاذ قراراتٍ بخصوص القيم المُدخَلة. ففي معالجة اللغة، يمكن أن تكون المدخلات أصواتًا، في حالة معالجة الصوت، أو أحرُفًا، في حالة معالجة الكلمة، أو جُمَلًا، في حالة معالجة النص، ويبقى السؤال هنا هو كيف يُمكِن الوصولُ لهذه الدوالِّ.

من الطرق المُستخدمة، والتي بدأ يقلُّ استخدامهـــا حاليًّا، النُّظُم الخـــبيرة (Expert Systems) (Giarratano & Riley, 2004). وتُبنى هذه النُّظُم ببرمجة قواعد المعالجة (والتي تكون عادةً جُمَلًا منطقية بصيغة إذا–فإن) وتكون الدالَّة في هذه الحالة هي هذه القواعد. فمثلًا في تطبيقات التحليل الصرفي، يقوم المطورون ببرمجة قواعد التحليل الصرفي للغة يدوياً، كذلك الحال في التحليل التركيبي، حيث تُبرمج قواعد النحو كاملة، وتكون مجموعة هذه القواعد هي المُكوِّن لدالَّة المعالجة. وتواجه هذه الطريقة صعوبات كثيرة، من أهمِّها كثرة قواعد التحليل وتعقيدها، إضافةً إلى وجود حالات كثيرة في اللغة مثل بعض الأسماء التي تحتاج إلى مُعالجة خاصَّة. فعلى سبيل المثال، هنالك نوع من التحليل يندرج تحت التحليل الصرفي–التركيبي (Morpho–syntactic Analysis). وإحدى المهام في هذا التحليل ما يعرف بتقطيع الكلمة (Word Segmentation)، وفيه تُفصل الأجزاء التي لها دور تركيبي، أي أنها تأخذ أحد أجزاء الكلام (Part of Speech)، عمَّا قبلها أو بعدها. مثال ذلك كلمة «ويذهبون». فإنه في هذه الكلمة تفصل الواو في بداية الكلمة «و» والتي تأخذ أحد أجزاء الكلام (وهو هنا رابط كونها عطفًا)، وكذلك تفصل الواو والنون «ون» في نهاية الكلمة والتي تأخذ أيضًا جزءًا من أجزاء الكلام (وهو هنا اسم بحكم عملها كفاعل)، بينما لا تفصل الياء «يـ» في «يذهبـ» لأنه ليس لها دور تركيبي وليس لها جزء من أجزاء الكلام، على الأقل في قائمة أجزاء الكلام المعتمدة لدى اتحاد البيانات اللغوية. يمكن ملاحظة أن

---

١ – ليس شرطا أن تكون الدالة كمية (تعالج أرقامًا). فقد تكون دالة منطقية أو غير ذلك. المهم هو أن تأخذ تعريف الدالة وهي أن تحول كل مدخل إلى مخرج واحد فقط.

٢ – سنشير دائما بالحرف الإنجليزي العريض لمتغير يُمثل مجموعة من القيم (مجموعة أو متجهًا)، بينما سنشير بالحرف العادي إلى متغير يأخذ قيمة واحدة.

هذا التحليل مختلف عن التحليل الصرفي الذي يهدف لاستخراج جذر الكلمة، إذ إنه في ذلك التحليل تفصل الياء أيضًا. أما في هذا التحليل فيكون الناتج «و- -يذهب- -ون». خُذْ على سبيل المثال أيضاً كلمة «واهم»، من الوَهْم، وهَبْ أن قواعد التحليل بُرمجت لفصل الواو في بداية الكلمة عمَّا بعدها، فإن هذه القاعدة لن تستطيع التفريق بين الواو التي من أصل الكلمة وبين واو العطف، فتقوم بفصل الواو في «واهم» لتُنتج «و- -اهم». هذا في الصفات والأحوال وأسماء الأجناس، والأمر أكثر تعقيدًا في أسماء الأعلام (Named Entities)، وهي أسماء الأشخاص والأماكن والمُنظَّمات. فلو وردت أيضا كلمة «الوليد» كاسم شخص في أحد السياقات، فقد تُحلل أيضًا وفق قاعدة فصل الألف واللام «الـ» لتكون «الـ- -وليد»، بينما من المُفترض ألَّا تُفصَل هنا كون الكلمة في هذا السياق اسمَ عَلَم وليست صفة. لذا فإننا في برمجة قواعد المعالجة نحتاج أن نأخذ كل هذه الاعتبارات في الحُسبان، وهذا أمر صعب جدًّا لعدم محدودية الكلمات، والتي تعتبر لامنتهية، إذ تدخل للغة كلمات جديدة بشكل مستمر، ولاختلاف السياقات التي تُستخدم فيها الكلمة الواحدة، والتي ربما يختلف تحليل الكلمة بناءً عليها. ومن الصعوبات أيضًا عدم وجود قواعد تحليل معروفة يُمكِن برمجتُها في كثير من التطبيقات، ففي كثيرٍ من تطبيقات تحليل الخطاب مثلًا، لا يوجد قواعد ثابتة معروفة مُتفق عليها يمكن برمجتها لتُمثل دالَّة التحليل، فلا يوجد قواعد ثابتة للتلخيص أو قراءة أسلوب الكتابة للتعرف على المُؤلف.

٣,٢ تعلم الآلة

ولتجاوز الصعوبات التي تواجه استخدام الأنظمة الخبيرة، يتوجَّه كثيرٌ من المطورين والباحثين إلى استخدام خوارزميات تعلُّم الآلة (Machine Learning)، والتي تهدف إلى مُحاكاة التعلم البشري. ففي حالة معالجة اللغة، فإن هذه الخوارزميات تُحاكي عمليات تعلم اللغة التي أشرنا إليها في بداية هذا الفصل، إذ تهدف إلى استقراء دوالّ معالجة اللغة من خلال الأمثلة التي تعطى لها. ففي حالة التحليل الصرفي مثلًا، تعطى هذه الخوارزميات مجموعة من الكلمات المُحلَّلة صرفيًّا، لتقوم هذه الخوارزميات باستقراء دالة[1] f تحاكي دالة التحليل الصرفي التي يستخدمها الإنسان. ونقول هنا «تحاكي» لأننا

١- الإشارة فوق الدالة ترمز إلى أنها دالة مُقدرة، وليست هي الدالة الحقيقة.

لا نعرف على سبيل اليقين كيف يقوم الإنسان بهذه العمليات. وللحصول على هذه الدالة، تقوم الخوارزميات عادةً بإجراءٍ يُسمَّى البحث والتقييم (Search and Score). ففي عملية البحث، تقوم الخوارزمية بالبحث في فضاء يُشار إليه بفضاء البحث (Search Space)، والذي يحتوي مجموعة من الدوال، عن دالة يمكن استخدامها، وأثناء الانتقال بين الدوال في هذا الفضاء، تُقيم كل دالة يُوصل إليها لفحص كفاءتها. يوجد العديد من الطرق المستخدمة في تقييم الدوال، والتي لا تتسع المساحة هنا لشرحها، لكن العامل المشترك لهذه الطرق هو الأخذ في الاعتبار كمية الأخطاء في النتائج التي تعطيها الدالة. لذا تهدف عملية البحث والتقييم إلى الوصول لدالة تعطي الكمية الأقل من الأخطاء على اللغة مطلقًا. ولكن المشكل هنا يكمن في صعوبة إثبات أن دالة ما لها الكمية الأقل من الأخطاء مطلقًا. إذا إنه حتى وإن أعطت دالة ما الكمية الأقلّ من الأخطاء على مجموعة من الظواهر، فإنها قد لا تعطي الكمية الأقل من الأخطاء على مجموعةٍ أخرى. وحيث إن إثبات أن دالة ما تعطي الكمية الأقل من الأخطاء مطلقًا يتطلب تجريب جميع الدوال وإظهار نتائجها على اللغة كاملة، والتي تحتوي على عدد لامنتهٍ من الظواهر. وهـذا يتطلب البحث والتقييم في فضاء مطلق لامنتهٍ من الدوال مع تقييم كل دالة في هذا الفضاء على اللغة كاملة، وهذه عملية غير منضبطة وغير قابلة للتطبيق. لذا فإنه في عملية البـحث والتقييم، يُحصر فضـاء البحث وذلك بوضع افتراضات مسبـقة (Assumptions) عن نوعيـة الدالة وشكلـها وصيغتها، وهذا ما يُعرَف بالانحياز الاستقرائي (Inductive Bias) أو انحياز التعلم (Learning Bias) (Mitchell, 1997). مع العلم أنه حتى بعد وضع هذه الافتراضات، فإن فضاء البحث قد يبقى لامنتهيًا ولكنه فضاء فرعي من الفضاء المطلق ومنضبط ومحصور في دوال مُعَرَّفَةٍ. وبدلًا من البحث عن الدالة التي تعطي العدد الأقل من الأخطاء مطلقًا، يُبحث عن دالة تعطي عددًا قليلًا من الأخطاء، لا يتجاوز حدًّا مُعيَّنًا، على مجموعة كبيرة من الظواهر اللغوية. والاختلاف في طرق تقييم الدوال يرجع إلى الاختلاف في كيفية تحديد هذا الحدِّ، وفي كيفية استخدام عدد الأخطاء في تقييم الدالة. لذا فإن استخدام الانحياز الاستقرائي هو ما يُفسر وجود عدد كبير من خوارزميات التعلم.

وسأُوضح هنا فكرة الانحياز الاستقرائي بمثالٍ[1]. لو أردنا تصميم خوارزمية لتعلم دالة لمهمة تقطيع الكلمة، والتي شرحناها في الجزء السابق، فيمكن بدايةً وضع الافتراضات التالية:

- المُدخلات إلى الدالة هي أرقام من مجموعة الاعداد الطبيعية تُمثل الأحرف الهجائية تسلسليًّا، أي أن الحرف «أ» له القيمة ١ والحرف «ي» له القيمة ٢٨.
- عدد المُدخلات إلى الدالة خمسة مدخلات، إحداها يُمثل الحرف الذي نريد أن نقرر بشأنه ما إذا كان يجب أن يُفصل عمَّا بعده أم لا، والأحرُف الأربعة الأخرى هي الحرفان اللذان قبله، والحرفان اللذان بعده. فلو أخذنا كلمة «يذهبون» وكُنَّا نريد معالجة الحرف «ب»، ستكون المدخلات (25 , 27 , 9 , 26 , 2) ، حيث إن الرقم الأول هو ممثل الحرف الذي تحت المعالجة، والأرقام الأخرى تمثل الحرفين الذين قبله، والحرفين الذين بعده تسلسليًّا.
- مُخرج الدالة يجب أن يكون عددًا حقيقيًّا أكبر من 0، إذا كان ما بعد الحرف الذي تحت المعالجة يجب أن يُفصل، أو أصغر من 0 إذا كان ما بعده يجب ألَّا يُفصَل. ففي المثال الذي في النقطة السابقة، من المُفترض أن تكون قيمة الدالة للمُدخَلات $\hat{f}(2,9,26,27,25) > 0$ لأن ما بعد الباء يجب أن يفصل بحيث تكون الكلمة «يذهب- -ون»، أما إذا كنا نُعالج الحرف «هـ» فتكون المدخلات وقيمة الدالة كالتالي $\hat{f}(26,28,9,2,27) < 0$، لأن ما بعد «هـ» يجب ألا يفصل.
- الدالة خَطيَّة، أي أنها تأخذ الصيغة التالية:

$$\hat{f}\big(\emptyset(\boldsymbol{x})\big) = \boldsymbol{w} \,.\, \emptyset(\boldsymbol{x}) + b$$

- عدد الأخطاء التي تعطيها الدالة يجب ألَّا تزيد نسبته عن ٨٪ من عدد الظواهر التي تقوم بمعالجتها.

---

١ – هذا المثال توضيحي فقط، ولا يهدف إلى شرح الطريقة المُثلى لحل مهمة تقطيع الكلمة.

هذه الافتراضات هي الانحياز الذي وضعناه لخوارزمية التعلم والتي تحصر البحث في فضاء الدوال الخطية فقط وفق القيود الأخرى المُوضَّحة في النقاط أعلاه. ففي هذه الحالة، ستكون عملية البحث مُقتصرةً على إيجاد قيم للمُتغيِّرات $\boldsymbol{w}$ و b والتي تجعل الدالة تعطينا عددًا قليلًا من الأخطاء في تحديد الحرف الذي يجب فصل ما بعده أم لا. يمكن ملاحظة أن فضاء البحث هنا يبقى لامنتهيًا أيضًا حتى بعد وضع هذه الافتراضات، إذ إن هنالك عددًا لامنتهيًا من القِيَم التي من الممكن أن تأخذها $\boldsymbol{w}$ وb[1]، وهو فضاء الأعداد الحقيقية، ولكنه أصغر بكثير من الفضاء المطلق الذي يحتوي على جميع الدوال. كما يمكن ملاحظة أن الدالة لا تأخذ المدخلات الرئيسية، وهي الأحرف، وإنما تأخذ مدخلات أخرى تمثل هذه الأحرف، وهي الأرقام المقابلة لها. لذا فإننا في كثيرٍ من الخوارزميات، نحتاج إلى دوال مُساعدةٍ تقوم بتحويل المدخلات الرئيسية إلى مدخلات أخرى تعمل عليها الدالة التي نبحث عنها. تُسمَّى هذه العملية بتحويل الخصائص (Feature Transformation). وفي المثال أعلاه، فإن الدالة $\varnothing$، هي التي تقوم بهذه المهمة.

وخوارزميات التعلم وطرق العمل عليها مُتعددة، إذ لا يمكن حصرها في فصل ولا حتى في كتاب واحد. ومن أبرز هذه الطرق، الشبكات العصبية (Neural Networks)، وهي نماذج رياضية تُبنى لتحاكي النظام العصبي للإنسان. إذ إن كل خلية عصبية تمثل دالة وتكون الشبكة بكاملها دالة مُركَّبة من مجموعة الدوال الأساسية التي تمثلها الخلايا العصبية. ومن أبرز الطرق الأخرى أيضا الطرق الاحتمالية، وهي ما سنتطرق له في الجزء التالي، وكذلك التعرف النمطي في الفضاء الدلالي، وهي ما سنختتم به هذا الفصل.

٣-الطرق الاحتمالية في تعلم الآلة

أشرنا في حديثنا عن صعوبة استخدام النظم الخبيرة إلى كثرة قواعد اللغة وتعقيدها، إضافة إلى أن بعض قواعد المعالجة غير معروفة على سبيل اليقين. وهذا ما يجعلنا نُضطرُّ إلى معالجة كثير من التراكيب اللغوية مع عدم اليقين (Uncertainty) بصحة المعالجة. لهذا، فإننا في هذه الحالة، نحتاج أن نصل إلى المعالجة الأقرب للصحة. ولكي نستطيع

١- ومع ذلك، فإنه يمكن الوصول للدالة التي تحقق الشروط المعطاة إن كانت موجودة وذلك باستخدام طرق رياضية مشهورة (Boyd & Vandenberghe, 2004).

تحديد القُرب والبُعد من صحة المعالجة، فإننا نحتاج إلى وضع مِعيار كَمِّيٍّ لدرجة الشك. فلو أخذنا المثال الذي ذكرناه في الغموض التركيبي، وهو تحديد قراءة جملة «رأيت الرجل جالسًا»، فإن درجة الشك لدينا بأن المراد هو القراءة الشائعة:

{ (رأى ت) [(الـ رجل) (جالسا)] }

أقلُّ بكثير من درجة الشك بأن المُراد هو القراءة الشاذة:

{[(رأى ت) (الـ رجل)] (جالسا) }

ذلك لأن القراءة الأولى هي المَعنيَّة في الغالب، ولكن لا يمكن القول بأن القراءة الأولى هي المُرادة على سبيل اليقين.

ومن النظريات المُستخدمة لقياس درجة الشك بشكل كمي نظرية الاحتمالات (Probability Theory). ولنظرية الاحتمالات تأصيلٌ رياضيٌّ يمكن الرجوع إليه في (Casella & Berger, 2001). ليكن x و y حدثينِ ولتكن P دالة احتمالية، فإنه:

- إذا كان x حدثًا مستحيل الوقوع، مثلًا x هو «عدم وجود حرف عربي في كلمة عربية»، فإن $P(x) = 0$.
- إذا كان x حدثًا يقينيًّا، مثلًا x هي «وجود حرف عربي في كلمة عربية»، فإن $P(x) = 1$.
- إذا كان x حدثًا مشكوكًا في وقوعه، مثلًا x هو «وجود حرف الضاد في كلمة عربية»، فإن $0 < P(x) > 1$ بحيث إنه إذا كان y حدثًا آخَرَ، مثلًا y هو «وجود حرف الهاء في كلمة عربية»، فإنه إذا كان شكُّنا بوقوع x أقل منه بوقوع y فإن $P(x) \geq P(y)$.
- احتمال حدوث إحدى حدثين x أو y لا يقعانِ معًا، مثلًا x هو «أن تبدأ كلمة ما بحرف الواو» و y هو «أن تبدأ كلمة ما بحرف التاء»، فإن احتمال حدوث أيٍّ من الحدثين هو $P(x \vee y) = P(x) + P(y)$.
- يُعتبر الحدثانِ x و y مُستقلَّينِ (Independent) إذا كانت معرفتنا بوقوع أحدهما لا تُغير من شكِّنا بوقوع الآخر. مثال ذلك لو كان لدينا نصان مختلفان T_1 و T_2،

وكان x هو «وجود كلمة –الذكاء– في T_1» وyهو «وجود كلمة –الاصطناعي– في T_2»، ففي هذه الحالة يكون احتمال وقوع الحدث x بعد معرفتنا بوقوع الحدث y، ويرمز له بالرمز $P(x|y)$ هو نفسه احتمال وقوع x، أي $P(x|y) = P(x)$.

- يعتبر الحدثان x و y مُرتبطينِ (Associated) إذا كانت معرفتنا بوقوع أحدهما تُغير من شكنا بوقوع الآخر. مثال ذلك لو كان لدينا نص T، وكان x هو «وجود كلمة –الذكاء– في T» وyهو «وجود كلمة –الاصطناعي– في T»، ففي هذه الحالة فإن احتمال وقوع الحدث x بعد معرفتنا بوقوع الحدث y، يتغيَّر عنه قبل معرفتنا بوقوع y، أي $P(x|y) \neq P(x)$.
- احتمال وقوع الحدثين x و y مع بعضهما، ونرمز له بالرمز $P(x \wedge y)$، هو:

$$P(x \wedge y) = P(y) * P(x|y)$$

وضعنا القواعد أعلاه على حدثين فقط بهدف تبسيط الشرح، ولكن يمكن تعميمها بنفس المنهجية على أكثر من حدثين. تُستخدم القواعد أعلاه لمعالجة الظواهر اللغوية مع عدم اليقين (Reasoning under Uncertainty) وذلك باختيار الحدث الذي له الاحتمال الأكبر. فبالعودة إلى مثال الغموض التركيبي، فعند حساب احتمال القراءتين، فإننا سنختار القراءة الشائعة:

{ (رأى ت) [(الـ رجل) (جالسا)] }

لأن احتمال حدوثها أكبر من احتمال حدوث القراءة الشاذة.

ويأتي دور خوارزميات تعلم الآلة في استقراء دالة التوزيع الاحتمالي (Probability Distribution) من الأمثلة التي تُعطى لها، وذلك بتحديد الأحداث المُستقلَّة والمرتبطة في البيانات المُعطاة، وتُعرَف هذه المهمة بتعلم بنية التوزيع الاحتمالي (Structure Learning)، وكذلك تحديد القيم الاحتمالية للأحداث، وتعرف هذه المهمة بتقدير المُعطيات (Parameter Estimation). وفي كثير من الاحيان، تكون بنية التوزيع الاحتمالي معروفة، ويقتصر استخدام تعلم الآلة على استقراء القيم الاحتمالية للأحداث فقط. وفي الجزء التالي، نُلقي الضوء على أحد النماذج الاحتمالية المستخدمة بكثرة وهي نماذج ماركوف الخفية.

١ , ٣ نماذج ماركوف الخفية

نماذج ماركوف الخفية (Hidden Markov Models –HMMs) هي نماذج احتمالية تُستخدم لتمثيل توزيعات احتمالية لها الخصائص التالية (Rabiner & Juang, 1986; Cappé, Moulines, & Ryden, 2007):

- يوجد أحداث مُتسلسلة خفية (غير محسوسة)، وسنُطلق عليها هنا حالات (States)، بحيث تكون كل حالة مرتبطة بالحالة التي قبلها. فلو كانت s_t حالة وقعت في الزمن t، فإن الحالة s_{t+1} والتي تقع في الزمن الذي بعده مباشرة $t+1$ غير مستقلَّة عنها. أي أن:

$$P(s_{t+1}|s_t) \neq P(s_{t+1})$$

- يُوجد أحداث ظاهرة (محسوسة)، وسنطلق عليها هنا انبعاثات (Emissions)، بحيث يكون كل انبعاث e_t في زمن ما t مرتبط بالحالة الخفية في نفس الزمن S_t. أي أن:

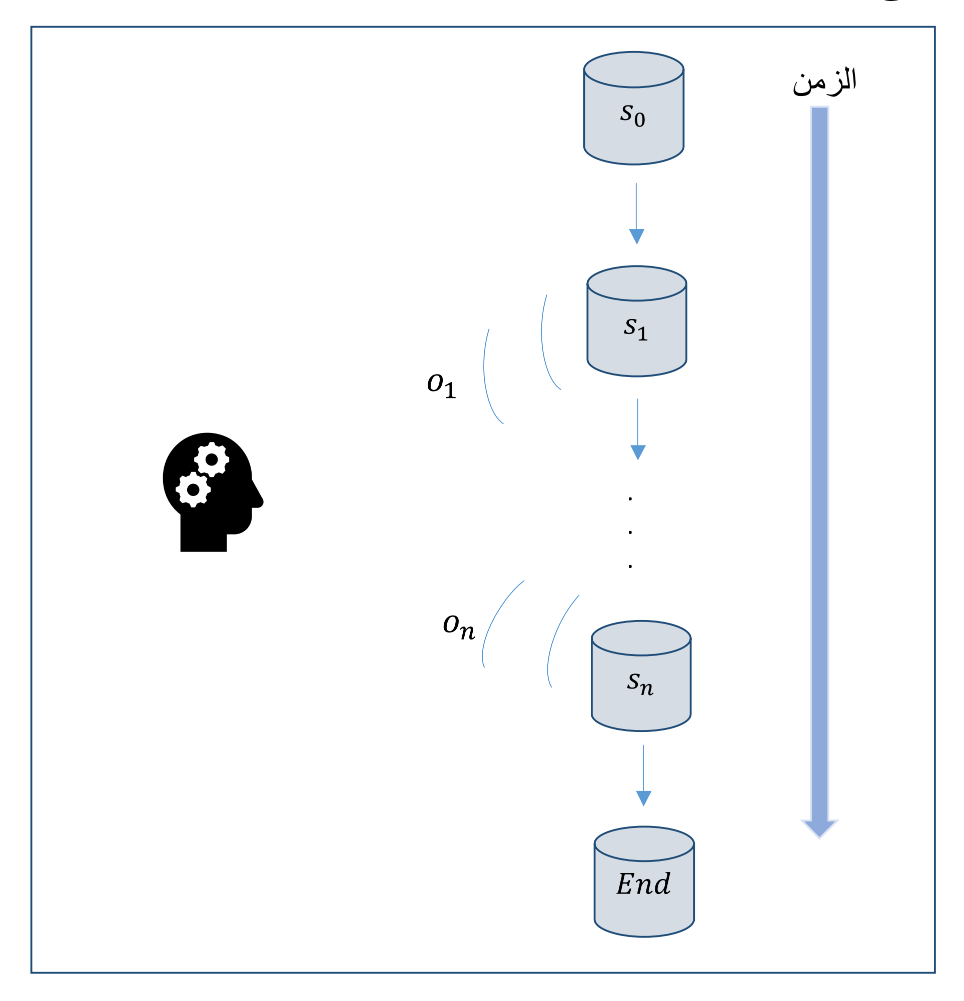
$$P(o_t|s_t) \neq P(o_t)$$

يُمكن تخيُّل نموذج ماركوف الخفي على أنه آلة لها هذا السلوك (الشكل ٢):

- عند تشغيلها تنتقل من حالة البداية S_0 إلى حالة ما S_1 لا نشعر بها وَفْق توزيع احتمالي $P(s_1|s_0)$. وبعد انتقالها إلى هذه الحالة تطلق انبعاثاً e_1 محسوسًا وَفْق توزيع احتمالي $P(o_1|s_1)$.
- ثم بعد ذلك تنتقل إلى حالة أخرى S_2 وفق توزيع احتمالي يعتمد على حالتها الآن وهو $P(s_2|s_1)$. ثم بعد انتقالها لهذه الحالة تُطلق انبعاثًا محسوسًا e_1 وفق توزيع احتمالي $P(o_2|s_2)$.
- وتستمرُّ في هذه العملية إلى أن تتوقف.

نلاحظ هنا أن انتقال الآلة من حالة إلى حالة، والانبعاثات التي تخرج من كل حالة، جميعها ليست يقينية. فلو كانت الآلة في الحالة الأولى S_1، فإنها يمكن أن تنتقل إلى أيِّ حالة من n من الحالات $s_2^1, s_2^2 \dots\dots, s_2^n$ في المرحلة التي تليها. وكذلك قد تُطلق أيَّ انبعاث من m من الانبعاثات $o_2^1, o_2^2 \dots\dots, o_2^m$. ولكن احتمالية الانتقال إلى الحالات

تختلف، فقد تكون احتمالية الانتقال إلى حالات معينة أعلى منها إلى الأخرى. فمثلا لو كانت:



الشكل ٢: رسم تخيُّلي لنماذج ماركوف الخفية.

$$P(s_2 = s_2^2|s_1) = 0.03 \text{ و } P(s_2 = s_2^1|s_1) = 0.6$$

فإن اعتقادنا بأن الحالة الثانية يمكن أن تكون s_2^1 أكبر منه بأن تكون الحالة s_2^2. والحال نفسه بالنسبة للانبعاثات. إذَنْ فإنه لا يمكن معرفة ما هي حالات الآلة بشكل يقيني، أما بالنسبة للانبعاثات فيمكن معرفة الانبعاثات التي ظهرت ولكن لا يمكن

معرفة الانبعاثات التي ستحدُثُ في المستقبل على سبيل اليقين. هذه العمليات تُعرَف في أدبيات الاحصاء والاحتمالات بالعمليات العشوائية[1] (Stochastic Processes).

هذا النموذج التخيُّلي يمكن تطبيقه على معالجة كثير من الظواهر اللغوية. ولنأخذ مثالًا وهو استنتاج أجزاء الكلام للكلمات الموجودة في الجُمَل (Part of Speech Tagging). فأجزاء الكلام تُعتبر حالات خفية غير موجودة في النص. حيث إن النص لا يحتوي سوى الكلمات والتي يمكن اعتبارها هنا انبعاثات تخرج من أجزاء الكلام. فالحالة الخفية «فعل»، على سبيل المثال، قد يخرج منها انبعاثات كثيرة وهي جميع الأفعال التي يعرفها الكاتب («ذهب»، «أكل»، «نام»، إلخ)، ولكن وفق احتمالات مختلفة، إذا أخذنا في الاعتبار أن بعض الأفعال أكثر شيوعًا من الأخرى. وكذلك فإن الانتقال من جزء كلام إلى جزء كلام آخر يتمُّ وَفْق قِيَم احتمالية مختلفة. فاحتمال الانتقال من حالة «فعل» إلى حالة «اسم» (أن يكون هنالك فاعل) ربما أعلى منه في الانتقال إلى حالة «حرف» (أن يكون الفاعل ضميرًا مسترًا). يمكن اعتبار أن الكاتب يمرُّ بعملية توليدية (Generative Process) أثناء كتابته للجمل بنفس العملية التي تمرُّ بها الآلة التي شرحناها في الأعلى. وهي أنه يبدأ الجملة بالانتقال من حالة البداية إلى واحدة من أجزاء الكلام. ثم يقوم بتوليد كلمةٍ بناءً على جزء الكلام الذي اختاره، وهي ما نراه في النص. ثم ينتقل إلى جزء كلام آخر بناءً على الجزء الحالي. وبعد انتقاله يقوم بتوليد كلمة أخرى بناءً على هذا الجزء الذي انتقل إليه. وهكذا حتى ينتهي من توليد الجملة كاملة. يمكن استنتاج أجزاء الكلام للكلمات في جملة ما، وذلك باختيار سلسلة أجزاء الكلام $s_1, s_2, \ldots, s_n$ التي لها القيمة الاحتمالية الأعلى وَفْق الدالة الاحتمالية أدناه، والتي تحسب احتمال مرور الكاتب بسلسلة من أجزاء الكلام أثناء توليده لجملة مُكوَّنة من الكلمات $o_1, o_2, \ldots, o_n$:

$$P(s_1 \wedge s_2 \wedge \ldots \wedge s_n \wedge o_1 \wedge o_2 \wedge \ldots \wedge o_n)$$

١ - يُطلق عليها «عشوائية» مجازًا، لكنها كما هو مُلاحظ ليست عشوائية بشكل مطلق، إذ إن احتمالات الأحداث مختلفة.

٢,٣ التعرف النمطي في الفضاء الدلالي

يُعرف التعرُّف النمطي (Pattern Recognition) بشكل مختصر على أنه التعرُّف على الظواهر من خلال بعض الأنماط التي تُصاحبها. وفي معالجة اللغة، هنالك العديد من الأنماط التي تُصاحب بعض الظواهر اللغوية والتي من الممكن استخدامها للتعرف على وجود هذه الظواهر. فعلى سبيل المثال، في تطبيقات التعرف على المُؤلِّف، يمكن تحليل مقالات عدَّة للمؤلف للتعرُّف على بعض الأنماط التي يستخدمها في أسلوب كتابته (الكلمات، تراكيب الجمل، طول الجمل، استخدام الروابط، الإملاء، وغيرها)، ومن ثَمَّ يُبحث عن هذه الأنماط في مقالةٍ مجهولةِ المُؤلف لمعرفة إمكانية أن تكون هذه المقالة قد كُتبت بواسطته أم لا. قد تكون الأنماط التي يُبحث عنها غير معروفة، وفي هذه الحالة تُستخدم خوارزميات تعلم الآلة لتعلمها من مجموعة كبيرة من النصوص، وقد تكون هذه الأنماط معروفة، وفي هذه الحالة تُبرمج دالة التعرف عليها مباشرة. وهذا الجزء يهدف إلى إعطاء فكرة مُبسَّطة عن إحدى أهمِّ طرق التعرُّف النمطي المستخدمة في المعالجة الدلالية والتي تُعرَف بتحليل الدلالة الكامنة ((Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998))، والتي لها العديد من الاستخدامات من أهمها معرفة الكلمات والسياقات المرتبطة دلاليًّا.

إحدى النظريات المطروحة في علم الدلالة تنصُّ على أن:

- المترادفات أو الكلمات التي لها معانٍ متقاربة (مثلًا: «عربة:سيارة»، «طائرة:سيارة»، «سيارة:شاحنة»)، تتواجد عادة في سياقات متشابهة، أي أنها تُحاط عادة بنفس المجموعة من الكلمات.

فعلى سبيل المثال، نرى أن كلمتَيْ «طائرة» و «سيارة» قد تظهر في هذه الجمل:

«سافرتُ بالطائرة أنا وصديقي».

«ركبتُ الطائرة حوالي الساعة الثانية عشرة».

«سافرتُ بالسيارة مع عائلتي».

«ركبتُ السيارة قبل قليل».

وهذه الجمل يوجد بينها تقاطع كبير في الكلمات التي تشير إلى نفس المفاهيم. لذا فإن هذه النظرية تُصنَّف تحت فرع في علم الدلالة يُسمَّى بالدلالة التوزيعية (Distributional Semantics) والذي يهتمُّ بدراسة المكونات الدلالية وتوزيعها في النصوص. يمكن اعتبار أن السياقات المتشابهة لكلمتين هي أحد الأنماط التي من خلالها يمكن معرفة تقارب هاتين الكلمتين دلاليًّا، أي أنهما مرتبطان بنفس المفهوم (Concept). كما أنه يُمكن اعتبار أن كمية الكلمات المتشابهة في السياقات مؤشراً على مدى ارتباط هذه السياقات دلاليًّا. وفي تحليل الدلالة الكامنة، تُمثل الكلمات والسياقات في مصفوفة بحيث تكون الكلمات هي صفوف المصفوفة، والأعمدة هي السياقات التي ظهرت فيها هذه الكلمات، كما هو موضح أدناه. ويطلق على هذه التمثيل الفضاء الدلالي :(Semantic Space)

سياق / كلمة	d_1	. . . .	d_m
w_1	٢		٠
w_2	١		١
. . . .	. . . .	. . . .	. . . .
w_n	٣		٠

تمثل خلايا المصفوفة عدد مرَّات ظهور الكلمة التي في الصف في السياق الذي في العمود[١]. فمثلاً، من خلال المصفوفة أعلاه، نجد أن الكلمة w_1 ظهرت مرتين في السياق d_1 ولم تظهر ولا مرَّة في السياق d_m، وهكذا. يُذكر أن السياق قد يكون مقالًا، أو مقطعًا، أو جملة، أو أيَّ جزء مُحدَّد من النص، وذلك بحسب الهدف من المعالجة. فلو أشرنا إلى المصفوفة أعلاه بالرمز $A_{n,m}$، والتي تحتوي n من الكلمات ظهرت في

١ – غالبًا لا يُستخدم عدد ظهور الكلمة، وإنما يستخدم وزن الكلمة في السياق. وهنالك طرق متعددة لوزن الكلمة يمكن الرجوع إليها في (Manning & Schütze, 1999)، ولكن استخدمنا هنا عدد ظهور الكلمة لتبسيط الشرح.

n من السياقات، فإن تحليل الدلالة الكامنة يقوم أولًا بتحليل هذه المصفوفة إلى ثلاث مصفوفات باستخدام تحليل رياضي يُعرَف بتفكيك القيمة المفردة (Singular Value Decomposition –SVD)[1] كالتالي:

$$A_{n,m} = U_{n,j} * S_{j,j} * (D_{m,j})^T$$ [2]

حيث إن المصفوفات الثلاث أعلاه هي عوامل (Factors) للمصفوفة الرئيسية $A_{n,m}$. ومن خلال هذا التحليل، تمثل المصفوفة $U_{n,j}$ الكلمات في المصفوفة الرئيسية ولكن في فضاء مختلف مُكوَّن من j من الأبعاد (Dimensions)، كالتالي:

بعد / كلمة	Dim_1	. . . .	Dim_j
w_1	-٠,١		٠,٢
w_2	-٠,٣		-٠,٥
. .	. .	. .	. .
w_n	٠,٤		٠,٣

ونفس الحال بالنسبة للمصفوفة $D_{m,j}$ التي تمثل السياقات في نفس الفضاء. هذا الفضاء الجديد المُكوَّن من j من الأبعاد يمثل فضاء المفاهيم وهو المشار إليه بكلمة «الكامن» في اسم طريقة التحليل، إذ إن هذا الفضاء غير ظاهر في الفضاء الدلالي الأصلي في المصفوفة الأصلية وإنما ظهر بعد تحليلها. في هذا الفضاء تتقارب الكلمات المرتبطة دلاليًّا، والتي تظهر عادة في نفس السياقات، وتكون قريبة من بعضها كما هو موضح في الشكل ٣، وكذلك الحال بالنسبة للسياقات التي تُشير إلى نفس المفاهيم. يمكن الحصول على الكلمات، أو السياقات، المتقاربة دلاليًّا في هذا الفضاء من خلال استخدام دالة لحساب بُعد المتجهات عن بعضها، حيث إنه كلما اقترب متجهان

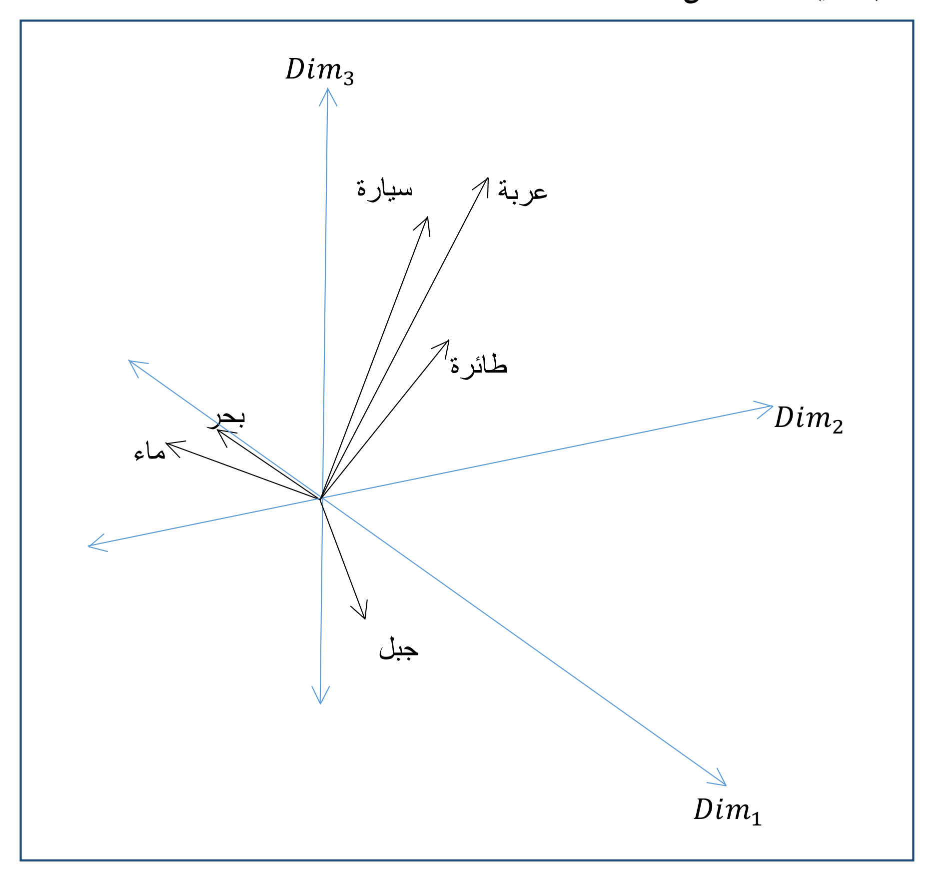
١– الترجمة هنا من الإنجليزية إلى العربية اجتهاد مني وقد لا تكون هي الترجمة المستخدمة في أدبيات الرياضيين العرب.

٢– D^T هو منقول المصفوفة D، والذي يعني تبديل الصفوف إلى أعمدة والأعمدة إلى صفوف، فتكون $D^T_{m,j} = D_{j,m}$

لكلمتين أو سياقين من بعضها في هذا الفضاء ازداد الارتباط الدلالي بينهما. ومن الدوال التي تستخدم لحساب قُرب متجهين دالة $cosine$ الشهيرة، والتي تقيس الزاوية بين متجهين $\boldsymbol{u}_1$ و $\boldsymbol{u}_2$ كالتالي:

$$cosine(\theta) = \frac{\boldsymbol{u}_1 . \boldsymbol{u}_2}{|\boldsymbol{u}_1| * |\boldsymbol{u}_2|}$$

حيث إن θ هي الزاوية بين المتجهين، و $\boldsymbol{u}_1 . \boldsymbol{u}_2$ هو الضرب النقطي (الجداء القياسي) لهما، و $|\boldsymbol{u}_i|$ هو طول المتجه $\boldsymbol{u}_i$. يُذكر أنه كلما قلَّت الزاوية بين المتجهين اقتربت قيمة الدالة من ١.



الشكل ٣: تقارب الكلمات التي لها نفس المفاهيم في الفضاء الدلالي الكامن. هنا افترضنا أن أبعاد الفضاء ثلاثة فقط ليسهل تمثيلها بصريًّا.

٤- الخاتمة

اللغة البشرية نظام مُعقَّد يمكن من خلاله إنتاجُ عدد لامنتهٍ من التراكيب اللغوية. ويهدف الباحثون في الذكاء الاصطناعي إلى فهم الإدراك البشري، ومحاولة محاكاته؛ وذلك بتطوير أنظمةٍ حاسوبيةٍ يُمكِن أن تعالج اللغة البشرية في مختلف مستوياتها. وتهدف هذه المعالجة في النهاية إلى تمكين الإنسان من التخاطب مع الآلة باستخدام اللغة التي يتخاطب بها مع أقرانه، ولكن هذا الهدف يواجه تحديات كثيرة اطلعنا في هذا الفصل على جزء منها. وهذه التحديات موجودة في جميع مستويات معالجة اللغة بدءاً من معالجة الصوت وحتى معالجة الخطاب. وتتركَّز أبرز الطرق المستخدمة حاليا في معالجة اللغات البشرية حول استخدام تعلُّم الآلة والتعرف النمطي. ويعكف الباحثون في مجالَيْ تعلُّم الآلة والتعرف النمطي على محاكاة تعلم الإنسان وطريقته في التعرف على الأنماط، ومن ثم محاكاة هذه الطرق وتطبيقها على مجالات عدة من ضمنها معالجة اللغات البشرية. وبالرغم من صعوبة الوصول إلى تطوير أنظمة حاسوبية يمكن أن تحاكي استخدام الإنسان للغة البشرية بشكل عام، إلا أن الباحثين نجحوا في تطوير العديد من الأنظمة التي تعالج مهامَّ محددة كفك الغموض التركيبي، أو التحليل الصرفي، أو تلك المتعلقة بالخطاب.

شكر وتقدير

الشُّكر لله سبحانه وتعالى أوَّلًا على تيسيره وإنعامه، ثم الشكر للوالدين الكريمين لدعمهما الدائم. بودِّي أن أتقدَّم بالشكر للأساتذة د. إبراهيم الخراشي، د. محمد الكنهل، ود. منصور الغامدي على جُهودهم المبكرة في دعم العمل البحثي على معالجة اللغة العربية في مدينة الملك عبدالعزيز للعلوم والتقنية. كما أتقدَّم بالشكر للأستاذة سارة العسكر على مُراجعتها اللغوية لهذا الفصل. أود أن أشكر جميع من عملت معهم في مدينة الملك عبدالعزيز للعلوم والتقنية على مشاريع في مجالَي تعلم الآلة ومعالجة اللغة العربية والتي كانت سبباً في تعلم الكثير.

أخيرًا وليس آخِرًا، أشكر زوجتي وأولادي على تفهُّمهم انشغالي المستمر خلال كتابة هذا الفصل.

المراجع

Alexander Clark. (2002). Unsupervised Language Acquisition: Theory and Practice. Essex: School of Cognitive and Computing Sciences, University of Sussex.

Christopher D. Manning، و Hinrich Schütze. (1999). Foundations of Statistical Natural Language Processing (1 الإصدار). MIT Press.

Daniel Jurafsky، و James H. Martin. (2008). Speech and Language Processing (2 الإصدار). Prentice Hall.

Daniel Marcu. (2000). The Theory and Practice of Discourse Parsing and Summarization. Daniel Marcu.

Eugene Charniak. (1997). Statistical parsing with a context-free grammar and word statistics. AAAI Press، 598-603 الصفحات.

George Casella، و Roger L. Berger. (2001). Statistical Inference (2 الإصدار). Duxbury Press.

Ion Androutsopoulos، و Prodromos Malakasiotis. (2010). A Survey of Paraphrasing and Textual Entailment Methods. Journal of Artificial Intelligence Research، 38(1)، 135-187.

James Allen. (1994). Natural Language Understanding (2 الإصدار). Pearson.

John J. McCarthy. (2018). Formal Problems in Semitic Phonology and Morphology (2 الإصدار). Routledge .

Jonathan Allen، M. Sharon Hunnicutt، Dennis H. Klatt، Robert C. Armstrong، و David B. Pisoni. (1987). From Text to Speech: The MITalk System. Cambridge University Press.

Joseph C. Giarratano، و Gary D. Riley. (2004). Expert Systems: Principles and Programming. Course Technology.

Lawrence R. Rabiner، و Biing-Hwang Juang. (1986). An Introduction to Hidden Markov Models. IEEE ASSP Magazine.

Mark Gales، و Steve Young. (2007). The Application of Hidden Markov Models in Speech Recognition (الإصدار Volume 1 Issue 3). Foundations and Trends in Signal Processing.

Olivier Cappé، Eric Moulines، و Tobias Ryden. (2007). Inference in Hidden Markov Models. Springer.

Sandra Kübler، و Emad Mohamed. (2011). Part of Speech Tagging for Arabic. Natural Language Engineering، 18(4)، 521-548.

Scott Deerwester، Susan T. Dumais، George W. Furnas، Thomas K. Landauer، و Richard Harshman. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science، 41(6)، 391--407.

Stephen Boyd، و Lieven Vandenberghe. (2004). Convex Optimization. Cambridge University Press.

Thomas K. Landauer، و Susan T. Dumais. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review، 104(2)، 211-240.

Thomas Landauer، Peter W. Foltz، و Darrell Laham. (1998). An Introduction to Latent Semantic Analysis. Discourse Processes، 2(3)، 259–284.

Tom M. Mitchell. (1997). Machine Learning (1 الإصدار). McGraw-Hill.

Wolfgang Sternefeld، و Wolfgang Sternefeld. (2013). Introduction to Semantics. De Gruyter Mouton.

الفصل الثاني

التعلم العميق وتطبيقاته في معالجة اللغة

د. فارس بن صالح القنيعير

ملخص الفصل

يعد تعلّم الآلة أحد المجالات الفرعية للذكاء الاصطناعي، ويهتم بصنع خوارزميات تتيح للحاسب التعلّم من البيانات للخروج بنماذج تفيد الكثير من التطبيقات، كمعالجة اللغات. إحدى مجموعات الخوارزميات التي انتشرت بشكل كبير في الفترة الأخيرة هي خوارزميات التعلّم العميق، التي هي امتداد لخوارزميات الشبكات العصبية. يرجع سبب انتشار استخدام التعلّم العميق إلى قدرتها على تعلّم نماذج بالغة التعقيد كان من الصعب تعلمها سابقاً، مما أتاح العديد من التطبيقات التي تعالج احتياجات واقعية، كرؤية الحاسب ومعالجة اللغات الطبيعية.

في هذا الفصل عرض موجز عن الشبكات العصبية والتعلم العميق. في البداية سيتم التحدث عن التسلسل التاريخي لتطور هذه الخوارزميات، ثم التطرق لأهم المعماريات المستخدمة، وفي النهاية عرض لبعض تطبيقاتها في معالجة اللغات الطبيعية، وذلك للخروج بفهم عام عن خوارزميات التعلم العميق وكيفية تطبيقها في مجال معالجة اللغات من دون الدخول في التفاصيل الدقيقة لكل خوارزمية، حتى يكون لدى الباحث تصور لما يمكن أن يقدمه التعلم العميق في المجالات المختلفة في معالجة اللغات الطبيعية.

د. فارس بن صالح القنيعير

حصل على درجة البكالوريوس في هندسة البرمجيات، ودرجة الماجستير في علوم الحاسب من جامعة الملك فهد للبترول والمعادن في المملكة العربية السعودية، ودرجة الدكتوراه في هندسة وتصميم النظم من جامعة واترلو في كندا. من اهتماماته البحثية: تعلم الآلة، تحليل الأنماط والتعرف عليها، ومعالجة الصور. وقد عمل على العديد من المشاريع البحثية مثل التعرف على لوحات السيارات السعودية، التعرف على الأشخاص عن طريق السمات الحيوية (القزحية والوجه)، التعرف على الأنسجة السرطانية وتصنيفها في صور الماموجرام، تقسيم وتحديد البروستاتا في صور الرنين المغناطيسي، التعرف على نوبات الصرع عن طريق إشارات الدماغ الكهربائية، وتصنيف النصوص والمشاعر في اللغة العربية.

١- مقدمة

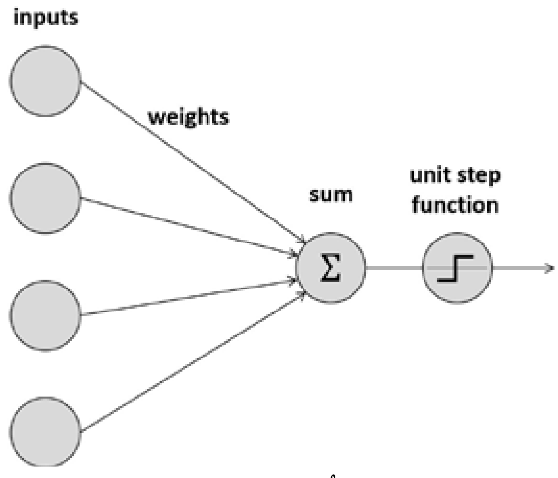
يعد تعلّم الآلة (Machine Learning) أحد المجالات الفرعية للذكاء الاصطناعي (Artificial Intelligence)، ويهتم بصنع خوارزميات تتيح للحاسب التعلّم من البيانات للخروج بنماذج تفيد الكثير من التطبيقات، كمعالجة اللغات. إحدى مجموعات الخوارزميات التي انتشرت بشكل كبير في الفترة الأخيرة هي خوارزميات التعلّم العميق (Deep Learning)، التي هي امتداد لخوارزميات الشبكات العصبية (Neural Networks). يرجع انتشار استخدام خوارزميات التعلّم العميق إلى قدرتها على تعلّم نماذج بالغة التعقيد كان من الصعب تعلمها سابقاً، مما أتاح العديد من التطبيقات التي تعالج احتياجات واقعية، كرؤية الحاسب (Computer Vision) ومعالجة اللغات الطبيعية (NLP).

سأتحدث في هذا الفصل عن الشبكات العصبية والتعلم العميق وتطبيقاتها في معالجة اللغات الطبيعية، والهدف الخروج بفهم عام عن خوارزميات التعلم العميق وكيفية تطبيقها في مجال معالجة اللغات من دون الدخول في التفاصيل الدقيقة لكل خوارزمية، حتى يكون لدى الباحث تصور لما يمكن أن يقدمه التعلم العميق في المجالات المختلفة في معالجة اللغات الطبيعية.

٢- تاريخ الشبكات العصبية والتعلّم العميق

قبل أن أتحدث عن التعلم العميق، يجدر أن أستعرض التسلسل التاريخي لخوارزميات الشبكات العصبية، وكيف تطورت إلى ما هي عليه الآن. ترجع البدايات لعام ١٩٤٣م، حيث قام وارن مكولش (Warren McCulloch) ووالتر بيتز (Walter Pitts) بوضع نموذج رياضي لكيفية عمل العقل وصنع نموذج للعصبونات (neurons)، التي تستخدم حتى الآن كمكون أساسي للشبكات العصبية (McCulloch & Walter, 1943). وفي عام ١٩٤٩م قام دونالد هيب (Donald Hebb) بشرح كيف أن الروابط بين العصبونات البيولوجية التي تستخدم سوياً تقوى مع كل استخدام (Hebb, 1949)، وهي توضح كيف يتم التعلم. في عام ١٩٥٨م قام فرانك روزنبلات (Frank Rosenblatt) بصنع جهاز المُستَقبِل (Perceptron) (Rosenblatt, 1958) ، وهو

محاكاة للنموذج الذي قام بوضعه كل من مكولش وبيتز عام ١٩٤٣. المُستَقبِل هو عبارة عن مصنّف خطي (linear classifier) يستقبل مُدخلات ويقوم بجمعها بشكل موزون حسب الأوزان ثم إخراج القيمة ٠ أو ١ بناءً على قيمة الحد (threshold). الشكل ١ يوضح خوارزمية المُستَقبِل.



الشكل ١: المُستَقبِل (Perceptron)

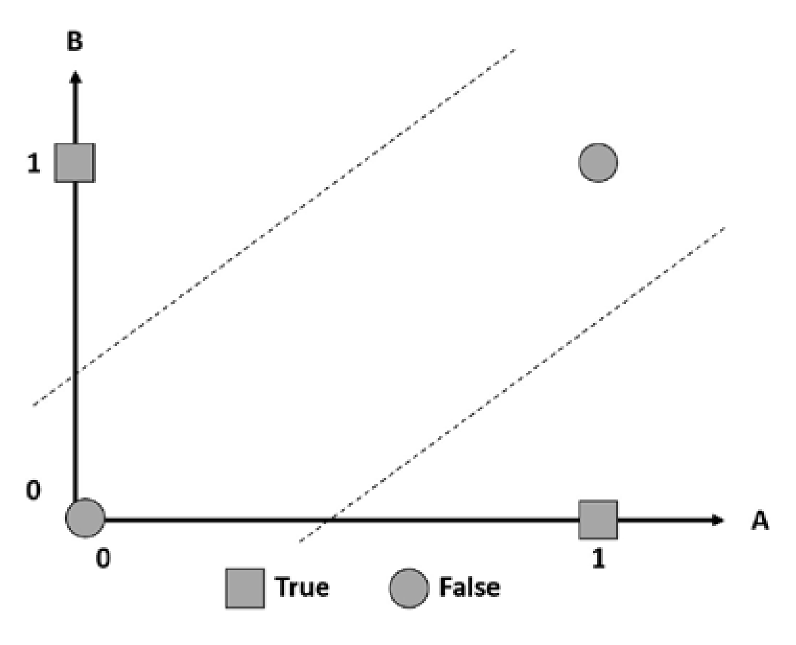
وفي هذه الأثناء كان كل من برنارد ويلدرو (Bernard Wildrow) ومارسيان هوف (Marcian Hoff) يعملان على نوع آخر من الشبكات العصبية تم نشره عام ١٩٦٠م أسمياه ADALINE (Widrow, 1960)، ثم طوراه عام ١٩٦٢م إلى MADALINE (Widrow, 1988 و Winter). وقد تم استخدام هذه الشبكات لإزالة الصدى من المكالمات الهاتفية، ولا تزال تستخدم تجارياً حتى الآن.

ما كان يعيب الشبكات العصبية في تلك الأثناء أنها كانت مصنّف خطّي لا تتمكن من تصنيف المشاكل اللّاخطّية (non-linear). وقد قام مارفن منسكي (Marvin Minsky) وسيمور بابرت (Seymour Papert) بتأليف كتاب عام ١٩٦٩م يوضح حدود خوارزمية المُستَقبِل perceptron ومشاكلها (Minsky و Papert, 1969). ولعل أشهر مثال ذكروه هو عجزها عن تصنيف XOR. الجدول التالي يوضح قيم XOR لمدخلين A وB:

A XOR B	B	A
0	0	0
0	1	1
1	0	1
1	1	0

جدول ١: قيم XOR

والشكل التالي يوضح أنه لا يمكن فصل القيم الناتجة باستخدام خط مستقيم، لأن المشكلة غير خطّية، لهذا لا يمكن استخدام خوارزمية المُستَقبِل لحل XOR.



الشكل ٢: تصنيف قيم XOR

بعد توضيح هذه المشاكل حدث جفاء كبير بين مجتمع الذكاء الاصطناعي وخوارزميات الشبكات العصبية، وانقطع دعم الأبحاث المتعلقة بها بشكل كبير. وقد استمر ذلك حتى بدايات عام ١٩٨٠م.

بدأ الاهتمام بالشبكات العصبية يعود بسبب عدة أحداث متوالية، بدءاً من مشاركة لجون هوبفيلد (John Hopfield) عام ١٩٨٢م في مؤتمر للأكاديمية الوطنية للعلوم (Hopfield، 1982)، حيث شرح الشبكة التي تحمل الآن اسم شبكة هوبفيد، وعاد بسببها الكثير من الباحثين إلى الشبكات العصبية. ثم تلاه إعلان اليابان عودتها لدعم الأبحاث المتعلقة بالشبكات العصبية، وتلا ذلك استحداث مؤتمرات سنوية ومجلات علمية متخصصة في الشبكات العصبية، كل ذلك زاد من زخم الدعم والنشر العلمي في هذا المجال. ولعل أهم الأمور التي أثَّرت في مسيرة الشبكات العصبية هما خوارزميتي الانتشار العكسي (backpropagation) والنزول الاشتقاقي (gradient descent). بالرغم من اقتراحها في الستينات، إلا أن خوارزمية الانتشار العكسي تم إعادة شرحها بشكل أوضح وإشهارها عن طريق كتاب (Learning Internal Representation by Propagation Error) الذي نشر عام ١٩٨٦م من تأليف روميلهارت (Rumelhart) وهينتون (Hinton) وويليامز (Williams) (Rumelhart، Hinton، و Williams، ١٩٨٦). وفي التسعينات وما بعدها تم اقتراح العديد من أنواع الشبكات العصبية التي لا تزال تستخدم حتى الآن، مثل LSTM وCNN، وسأتكلم عنهما لاحقاً في هذا الفصل.

كما ذكرت سابقاً، التعلم العميق هو فعلياً شبكات عصبية ولكن بطبقات أكثر. فتاريخ التعلم العميق مرتبط بشكل كبير بالشبكات العصبية. ولكن كان هناك مشاكل تعيق تدريب شبكات بهذا التعقيد، كقلة البيانات وضعف القدرة الحاسوبية وبعض المشاكل في الخوارزميات التي تم حلها تدريجياً. الموجة الثالثة في انتشار استخدام الشبكات العصبية هي عهد التعلم العميق، حيث بدأت على الأرجح عام ٢٠٠٦م ببحث منشور يشرح كيفية تدريب شبكات عميقة من نوع Deep Belief Networks (Hinton، Osindero، و Teh، ٢٠٠٦). ولكن الشهرة الحقيقية التي سببت انتشار استخدام التعلم العميق هو فوز خوارزمية تعلم عميق AlexNet (Krizhevsky، Sutskever، و Hinton، ٢٠١٢)) بتحدي Large Scale Visual Recognition Challenge (ILSVRC) عام ٢٠١٢م بالمركز الأول بفارق كبير جداً بين المركزين الأول والثاني. هذا التفوق الكبير فتح أعين الباحثين على القدرة الكبيرة للتعلم العميق في بعض المجالات كرؤية الحاسب ومعالجة اللغات الطبيعية. ولا يزال المجتمع البحثي نشط جداً في الأبحاث المتعلقة بالتعلم العميق وكيفية تطبيقه في مختلف المجالات.

٣- أسباب نجاح التعلم العميق

كما يتضح من تاريخ الشبكات العصبية، فالكثير من المفاهيم المستخدمة في التعلم العميق تم استخدامها منذ زمن بعيد. ولكن هناك عدة أسباب أدت إلى نجاح التعلم العميق لاحقاً، يمكن اختصارها في أربعة أسباب أساسية:

١- **البيانات الضخمة**: مع رخص وسائل التخزين وزيادة سعاتها، إضافة إلى سهولة تسجيل البيانات وتنوعها صار بالإمكان جمع بيانات ضخمة. أحد متطلبات تدريب نماذج التعلم العميق المعقدة هو توفر بيانات ضخمة يمكنها تعلم الملايين من الأوزان.

٢- **المعالجات الرسومية**: يتطلب تدريب الشبكات العميقة عمليات حسابية كثيرة جداً، حيث يتم تعلم ملايين الأوزان. باستخدام المعالجات الرسومية صار بالإمكان توزيع العمليات الحسابية بالتوازي (parallel)، مما ساهم في تسريع التدريب بشكل كبير.

٣- **تطور خوارزميات الشبكات العصبية**: مما ساهم في نجاح تدريب الشبكات العميقة حل بعض المشاكل كتلاشي المشتقة (vanishing gradient) وانفجار المشتقة (exploding gradient). وكذلك اقتراح استخدام دوال تفعيل جديدة مثل دالة ريلو، وغيرها من التطويرات العديدة.

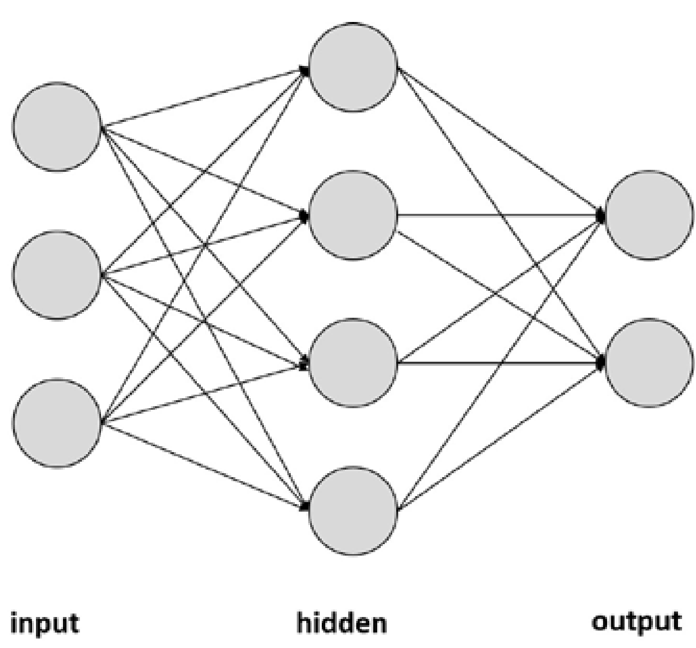
٤- **ثقافة المشاركة في مجتمعات الذكاء الاصطناعي وتعلم الآلة**: أحد العوامل المهمة في انتشار استخدام التعلم العميق هو ثقافة المشاركة في مجتمعات الذكاء الاصطناعي. وثقافة المشاركة تتضمن نتائج الأبحاث والأوراق العلمية عن طريق نشرها مجاناً على مواقع مثل arxiv.org، مما يتيح للجميع الوصول لها بدون اشتراكات باهظة الثمن. بالإضافة إلى مشاركة الأكواد والبرامج والبيانات.

٤- الشبكات العصبية والتعلم العميق

تعد الشبكات العصبية من الخوارزميات المهمة في مجال تعلّم الآلة، وهي تتبع لمدرسة تسمى التشبيكية (connectiontist)، والتي استقت أفكارها من محاولة محاكاة الدماغ البشري وتشابك الأعصاب. فكما شرح دونالد هيب بأنه عند التعلم تقوى روابط

العصبونات التي تستخدم سوياً، وهي الفكرة الأساسية التي تقوم عليها الشبكات العصبية، حيث تسعى الخوارزمية أن تتعلم أوزان الروابط بين العصبونات. الشبكات العصبية تستطيع تعلم مشاكل غير خطية في غاية التعقيد.

تتكون الشبكات العصبية بشكل أساسي من عصبونات (neurons) وأوزان الروابط (weights) ودوال تفعيل (activation functions)، وكذلك من مدخلات (inputs) ومخرجات (outputs)، كما هو موضح في الشكل ٣. وهي تتكون غالباً من عدة طبقات (layers).



الشكل ٣: شبكة عصبية بسيطة

يتم حساب قيمة كل عصبون في الطبقات الداخلية عن طريق ضرب قيم الطبقة التي تسبقها بالأوزان وإضافة قيمة الانحياز b، ثم إدخال النتيجة إلى دالة التفعيل كما هو موضح في المعادلة:

$$h(X, W, b) = \emptyset(XW + b) = \emptyset(\sum_{i=1}^{n} x_i \cdot w_i + b_i)$$

حيث X مصفوفة تحتوي على قيم المدخلات أو الطبقة السابقة، W مصفوفة بقيم الأوزان، و b متجه بقيم الانحياز (bias)، و $\varnothing$ هي دالة التفعيل. يجدر التأكيد أن المعادلة السابقة تمثل حساب عصبون واحد فقط، ويجب أن تحسب لكل عصبون في كل طبقة.

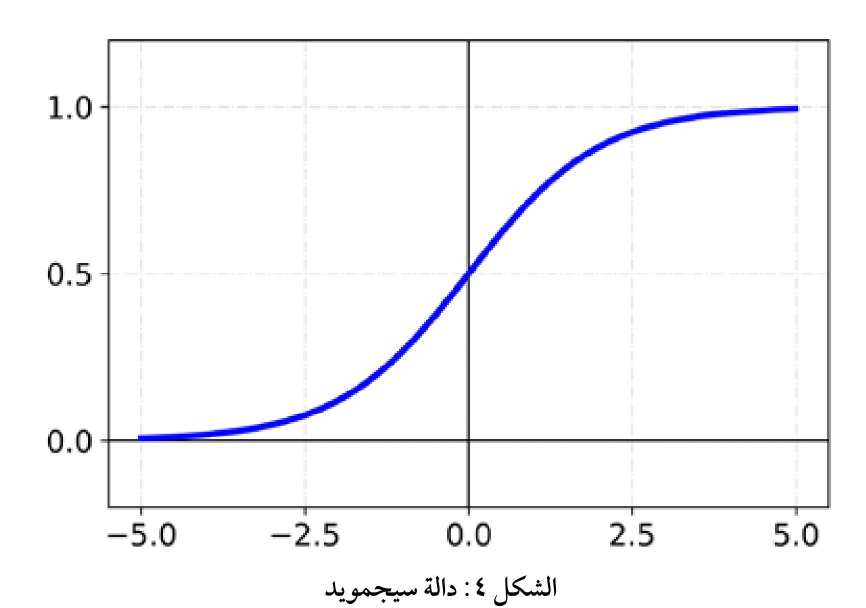
تعد دالة التفعيل من المكونات الأساسية للشبكات العصبية، فمن خلالها تكتسب قوتها في التصنيف غير الخطي. هناك عدة أنواع لدالة التفعيل، سابقاً كان الأكثر استخداماً هما دالتي سيجمويد (sigmoid):

$$\emptyset(z) = \frac{1}{1 - e^{-z}}$$

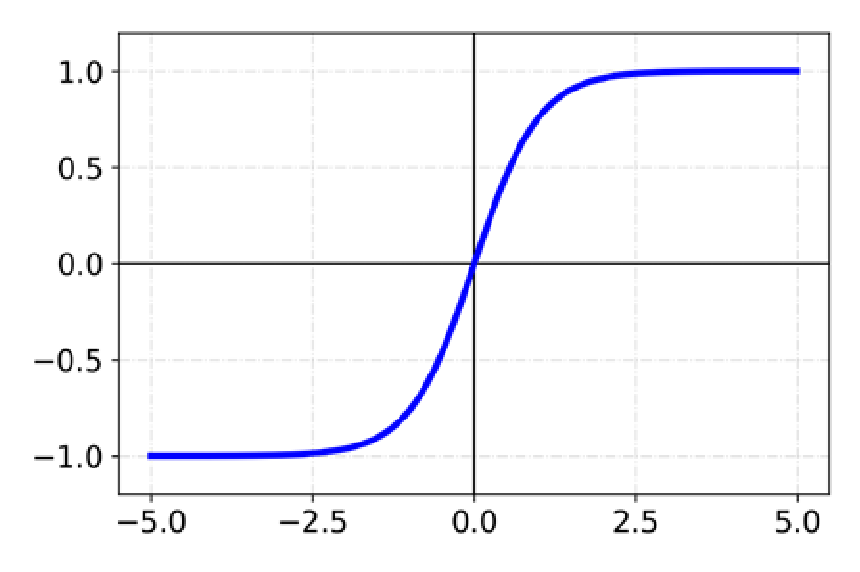
ودالة الظل الزائدي (Tanh):

$$\emptyset(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

الموضحين في الشكلين ٤ و ٥.



الشكل ٤: دالة سيجمويد

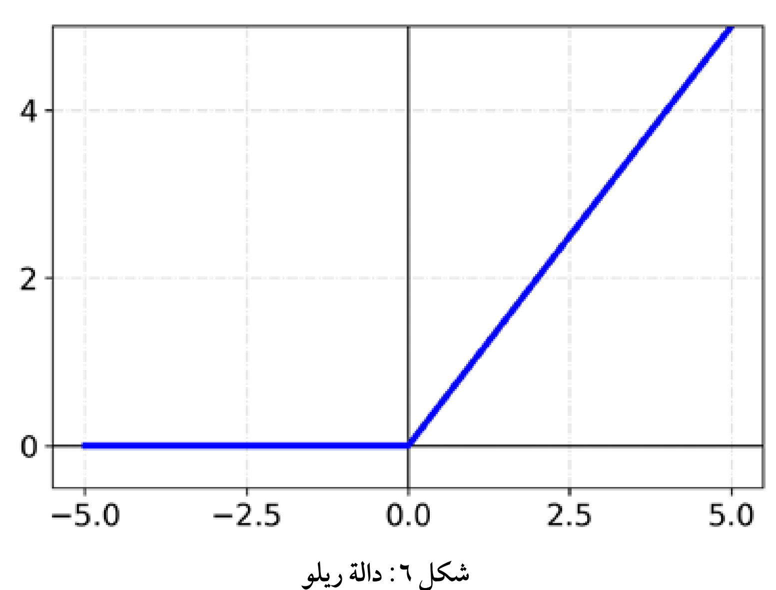


الشكل ٥ : دالة Tanh

ولكن مؤخراً تم التوجه إلى دوال أخرى أكثر فعالية للتعلم العميق، أشهرها وأكثرها استخداماً هي دالة ريلو (ReLU):

$$\emptyset(z) = \max\,(0, z)$$

الموضحة في الشكل ٦.



شكل ٦ : دالة ريلو

١, ٤ تدريب الشبكات العصبية

الخوارزمية الأشهر استخداماً لتدريب الشبكات العصبية هي خوارزمية الانتشار العكسي (backpropagation). وفيها يتم استهلال الأوزان عشوائياً في البداية ثم حساب المخرجات كما تم شرحه لكل عصبون في كل طبقة. بعد ذلك يتم حساب دالة التكلفة (cost function) التي توضح مقدار الخطأ في المخرجات. هناك عدة دوال لحساب التكلفة، مثل دالة Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

حيث y هي القيم الحقيقية لعينات التدريب، و$\hat{y}$ هي قيمة المخرجات من الشبكة العصبية.

بناءً على ذلك يتم تحديث الأوزان لتقليل دالة التكلفة بشكل تكراري عن طريق حساب النزول الاشتقاقي (gradient descent) واستخدامه لتحديث الأوزان، حتى الوصول إلى نتيجة مقبولة.

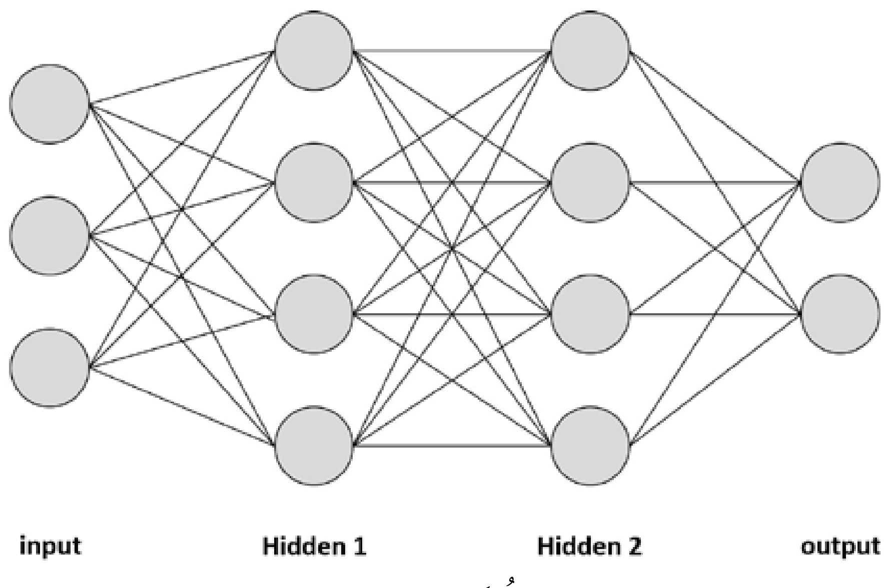
٥- معماريات الشبكات

تتنوع معماريات الشبكات العصبية حسب التطبيق المطلوب، حيث كل معمارية لها خصائص لا تتوفر بغيرها. وسيتم التطرق هنا لثلاثة من أكثر المعماريات استخداماً. بشكل عام يتم إطلاق التعلم العميق على الشبكات العصبية ذات الطبقات الكثيرة، ولا يوجد رقم محدد متفق عليه لعدد الطبقات حتى نطلق على الشبكة شبكة عميقة، فبعضهم يعدها ١٠ وبعضهم أقل أو أكثر. وكلما زاد عدد الطبقات زادت إمكانية الشبكة لتمثيل وتعلم مفاهيم أعقد.

١, ٥ المُستَقبِل متعدد الطبقات (Multi-Layer Perceptron (MLP))

يعد المستقبل متعدد الطبقات أحد أشهر خوارزميات الشبكات العصبية، وهي النسخة التي يكون شرح الشبكات العصبية عليها في البداية غالباً، وقد تم شرحها في الفصل السابق، انظر الشكل ٧. ويتم تسمية طريقة تشابك طبقاتها بـ «الطبقات تامة

الاتصال» (fully connected layers)، حيث في الغالب يتصل كل عصبون في طبقة مع جميع العصبونات في الطبقات التي تسبقها.



الشكل ٧: شبكة المُستَقبِل متعدد الطبقات (MLP)

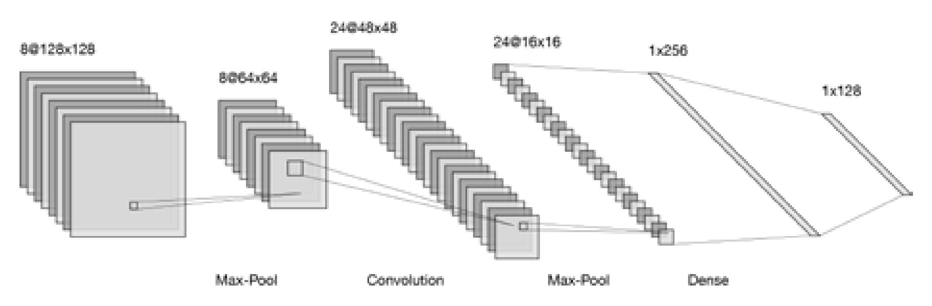
٢, ٥ الشبكات العصبية الترشيحية (Convolutional Neural Networks)

تتعامل الشبكات العصبية الترشيحية في الغالب مع مصفوفات ثنائية الأبعاد، والتي تكون على الأرجح صور. طريقة التعلم قريبة من الآلية التي تم شرحها هنا، ولكن الفرق في طريقة تمثيل الطبقات وتشابهها. في الشبكات الترشيحية بدلاً من تعلم الأوزان بين كل عصبون والمقابل له في الطبقة التي تليه، يتم تعلم عدة مرشحات (filters / kernals) يمكن تطبيقها على الصور ككل. بهذه الطريقة يتم تقليل عدد الأوزان التي يجب تعلمها بشكل كبير جداً، مما يسهم في تسريع عملية التعلم وتقليل فرط التخصيص (overfitting). هذه الخاصية يطلق عليها مشاركة المُدخلات (parameter sharing). هناك عدة أنواع للطبقات في الشبكات الترشيحية (أنظر الشكل ٨)، أهمها:

١- **طبقات الترشيح** (convolutional layer): وفيها يتم تطبيق المرشحات التي يتم تعلم أوزانها.

٢- **طبقات التقليص** (layer pooling): وفيها يتم تقليص حجم الصور، وقد يكون التقليص بالمعدل (mean pooling) أو بالقيمة الأكبر (max pooling).

٣- **الطبقات تامة الاتصال** (fully connected layers): وهي مثل التي تم شرحها في السابق، يتم تحويل المصفوفات ثنائية الأبعاد إلى متجه من بعد واحد. وقد يكون هناك أكثر من طبقة تامة الاتصال قبل طبقة المخرجات (output layer).



الشكل ٨: شبكة عصبية ترشيحية

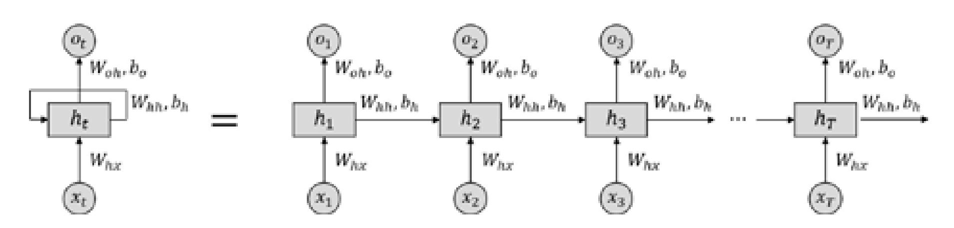
٣,٥ الشبكات العصبية التكرارية (Recurrent Neural Networks)

في أنواع الشبكات التي تم شرحها حتى الآن لا يؤخذ الزمن أو العلاقة بين سلاسل البيانات بالاعتبار. ولكن هناك العديد من التطبيقات التي يجب أن تأخذ في الحسبان علاقة البيانات بين بعضها في السياق الزمني، كالتعرف على الكلام (Speech Recognition)، أو المكاني كالتعرف على النصوص المطبوعة (Optical Character Recognition). في الشبكات التي تم شرحها حتى الآن المدخلات والمخرجات مستقلة عن بعضها، لذا يصعب أن نستخدم السياق الذي تأتي المدخلات فيها (كمكان الحرف في الكلمة أو الكلمة في الجملة). الشبكات العصبية التكرارية تحل هذه المشكلة عن طريق **تذكر** ما تم تعلمه من المدخلات السابقة، وبهذا يمكن تعلم الحالة الماضية واستخدامها مع المدخلات الحالية، أنظر الشكل ٩. ويتم حساب قيمة كل عصبون كالتالي:

$$h_t = \emptyset(W_{hx}x_t + W_{hh}h_{t-1} + b_h)$$

حيث h_t تأخذ قيمة المدخلات مضروبة بالأوزان الخاصة بها، وقيمة h_{t-1} مضروبة بالأوزان الخاصة بها، وt تشير إلى الترتيب.

من المشاكل التي تواجهها هذه الشبكات هي تلاشي المشتقة (vanishing gradient) وانفجار المشتقة (exploding gradient) في السلاسل الطويلة. وقد تم اقتراح عدة خوارزميات لحلها أشهرها LSTM وGRU، وسيتم شرحها الآن.



الشكل ٩: شبكة عصبية تكرارية

١,٣,٥ الذاكرة قصيرة المدى المطولة (Long Short-Term Memory (LSTM))

تم نشر خوارزمية الذاكرة قصيرة المدى المطولة (LSTM) في بحث لهوتشريتر (Hochreiter) وشميدهوبر (Schmidhuber) عام ١٩٩٧م (Hochreiter و Schmidhuber، ١٩٩٧)، وقد تم اقتراح العديد من التحسينات والأنواع المختلفة لها بعد ذلك. تحل LSTM المشاكل التي تواجهها شبكات RNN بحيث تقلل من حدة التغير في المشتقات مقارنة بـ RNN، كما أنها مصممة بحيث يكون لديها نوعين من الذاكرة، قصيرة المدى وطويلة المدى. الفارق الأساسي في LSTM هي آلية التذكر والنسيان، وقد تم تصميم الشبكة عن طريق استبدال الطبقات البسيطة بأخرى أكثر تعقيداً تتكون من عدة بوابات، يطلق على هذا النوع من الطبقات خلية (cell)، الشكل ١٠ يوضح شكل خلية LSTM. تحتوي LSTM على ثلاثة أنواع من البوابات: بوابة إدخال (input gate)، بوابة نسيان (forget gate)، وبوابة إخراج (output gate). ويتم حساب كل من الذاكرة (c) (cell state) والمخرجات/ أو الحالة (h) (hidden state) كالتالي:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$

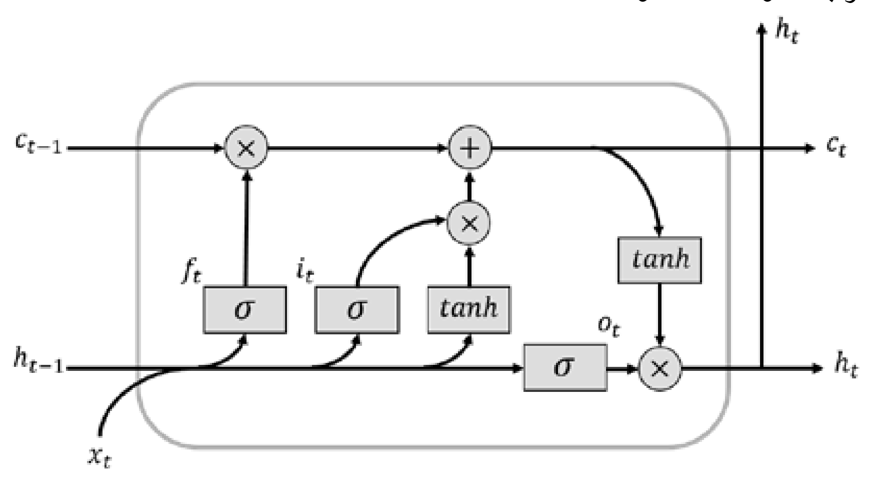
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$h_t = o_t \odot tanh(c_t)$$

حيث x_t المدخلات، f_t و i_t و o_t متجهات بوابات النسيان والإدخال والإخراج على التوالي، σ دالة التفعيل سيجمويد، WوU وb الأوزان لكل من البوابات السابق ذكرها والمدخلات والذاكرة والحالة السابقة، وt وحدة الزمن أو الخطوات، و $\odot$ ترمز لضرب مكونات المصفوفات (element-wise multiplication).



الشكل ١٠: خلية الذاكرة قصيرة المدى المطولة (LSTM)

٢,٣,٥ الوحدة التكرارية المُبَوَّبَة ((Gated Recurrent Unit (GRU))

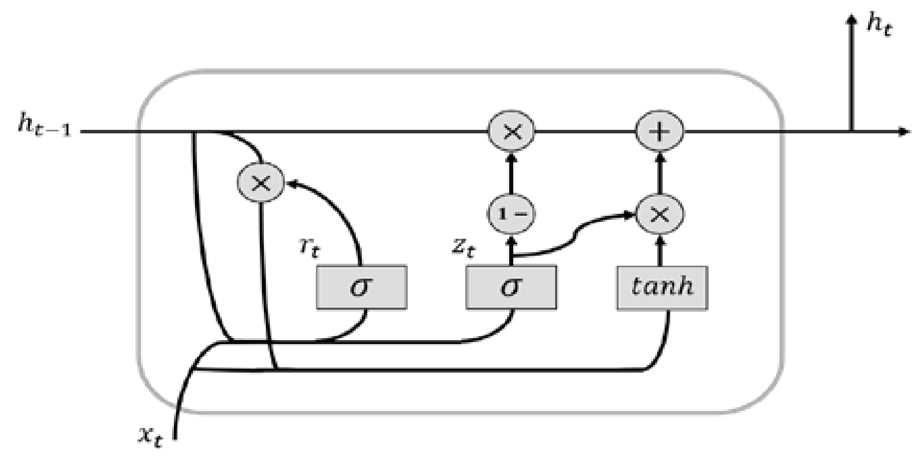
أحد الخوارزميات المشتقة من LSTM هي الوحدة التكرارية المبوبة (GRU) (Cho، Merrienboer، Bahdanau، و Bengio، ٢٠١٤)، وتعد تبسيطاً لها من عدة جهات. فمثلاً في GRU تم دمج بوابتي الإدخال والنسيان إلى بوابة واحدة اسمها بوابة التحديث (update gate)، كما تم دمج خلية ذاكرة الخلية وحالة الخلية (cell and hidden states). الشكل ١١ يوضح شكل خلية GRU. كما هو واضح فإنها أبسط من خلية LSTM، ومع هذا فإن الأداء بين GRU وLSTM متقارب جداً، مما أدى إلى تبنيها بشكل كبير نظراً لكفاءتها. المعادلات التالية تبين كيفية حساب المخرجات:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$h_t = (1 - z_t)\odot h_{t-1} + z_t \odot \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

حيث z_t متجه بوابة التحديث (update gate)، و r_t متجه بوابة إعادة التعيين (reset gate).



الشكل ١١: خلية الوحدة التكرارية المُبَوَّبَة (GRU)

٦- تطبيقات التعلم العميق في معالجة اللغة

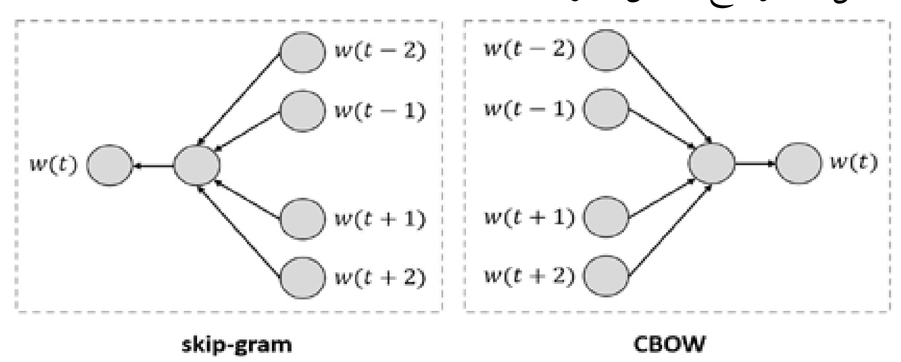
تتنوع فروع معالجة اللغات الطبيعية ((Natural Language Processing (NLP)) إلى العديد من المجالات والتطبيقات، ويرجع العمل عليها للعديد من العقود. سيكون التركيز في هذا الفصل على كيفية تطبيق التعلم العميق في هذه المجالات. وبما أن العشرات والمئات من الأبحاث قد نشرت لكل من هذه المجالات، فليس المجال هنا الحصر، ولكن لإعطاء فكرة عن المجالات المختلفة وكيفية استخدام التعلم العميق فيها، وسيتم شرح كل مجال وذكر مثال لحله باستخدام التعلم العميق.

١,٦ تضمين الكلمات (Words Embeddings)

تضمين الكلمات هو تمثيل للكلمات على شكل متجهات (vectors) محافظة لترابط الدلالات. من فوائد تضمين الكلمات أن طول المتجه الذي يمثل الكلمات أقل بكثير من عدد الكلمات المستخدمة. فمثلاً بعض التمثيلات التي كانت تستخدم مثل One Hot Encoding إذا كان لدينا ٥٠ ألف كلمة فسيتم تمثيل كل كلمة بمتجه طوله ٥٠ ألف، بحيث يكون كله أصفار ما عدا مكان الكلمة يكون واحد. بينما الكلمات المضمنة

يتم تمثيلها بمتجه طوله غالباً بين ١٠٠ و٣٠٠، ومكون من أرقام يتم تعلمها. إحدى خصائص هذه الخوارزميات هي المحافظة على المعنى الدلالي للكلمات، بحيث تكون الكلمات المتقاربة في المعنى قريبة من بعضها في فضاء المتجهات.

هناك عدة خوارزميات لتضمين الكلمات، أشهرها word2vec (Mikolov، Sutskever، Chen، Corrado، و Dean، ٢٠١٣)، Glove (Pennington، Socher، و Manning، ٢٠١٤)، و fasttext (Bojanowski، Grave، Joulin، و Mikolov، ٢٠١٧). هناك طريقتان لتعلم التضمين في word2vec، إما باستخدام continuous bag of words (CBOW)، أو skip-gram. الهدف في CBOW هو تعلم الكلمات المناسبة من السياق (context)، بينما في skip-gram هو تعلم السياق من الكلمات. الشكل ١٢ يوضح كلاً من الطريقتين.



الشكل ١٢: شبكتا تضمين الكلمات باستخدام CBOW و skip-gram

٢,٦ التعرف على المشاعر (Sentiment Analysis)

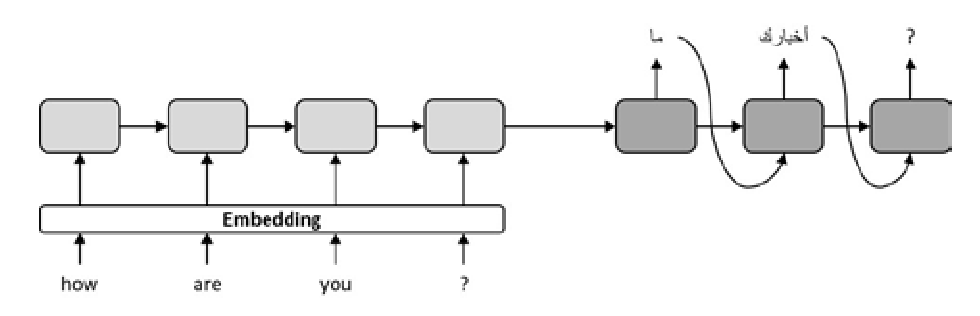
كثيراً ما تحتاج الجهات أن تعرف مشاعر العملاء عن الخدمات والمنتجات التي تقدمها، إحدى الطرق التي انتشر استخدامها مؤخراً استخدام خوارزميات التعرف على المشاعر لتحليل النصوص ومحاولة معرفة مشاعر الكاتب. يمكن تصنيف المشاعر إلى ثلاث أو خمس فئات، أو غيرها من التصنيفات. ويمكن تدريب نماذج التعرف على المشاعر لتعمل على النص كاملاً، أو على الفقرات كل على حدة.

غالب الطرق التي تستخدم التعلم العميق تبدأ بتحويل النص إلى تمثيل الكلمات المضمنة الذي تم شرحه قبل قليل. ولأن السياق وأخذ الجمل كاملة في الاعتبار مهم فالكثير يستخدم أحد الأشكال المختلفة من معمارية RNN، كـ LSTM أو GRU.

٦,٣ الترجمة الآلية (Machine Translation)

الترجمة من لغة إلى أخرى آلياً مجال خصب للأبحاث، وقد خرجت العديد من الأبحاث والخوارزميات التي تستخدم التعلم العميق لتعلم الترجمات. موقع ترجمة قوقل بدأ باستخدام التعلم العميق من عام ٢٠١٦م.

يتم في الترجمة الآلية (والتعرف على الكلام والتلخيص الآلي كما سيأتي) إدخال سلسلة من المدخلات وإخراج سلسلة من المخرجات، يطلق على الشبكات التي تستخدم لهذا النوع من التطبيقات Sequence-to-Sequence (Seq2Seq) (Sutskever، Vinyals، و Le، ٢٠١٤)، وفيها يتم استخدام شبكتين من نوع RNN بحيث يتم ترميز السلسلة الأولى (encoder) وفك الترميز للشكل المستهدف (decoder). في الترجمة الآلية تكون السلسلة الأولى اللغة المصدر والسلسلة الثانية اللغة المستهدفة. الشكل ١٣ يوضح شبكة Seq2Seq للترجمة الآلية، ويطلق عليها أيضاً encoder-decoder.

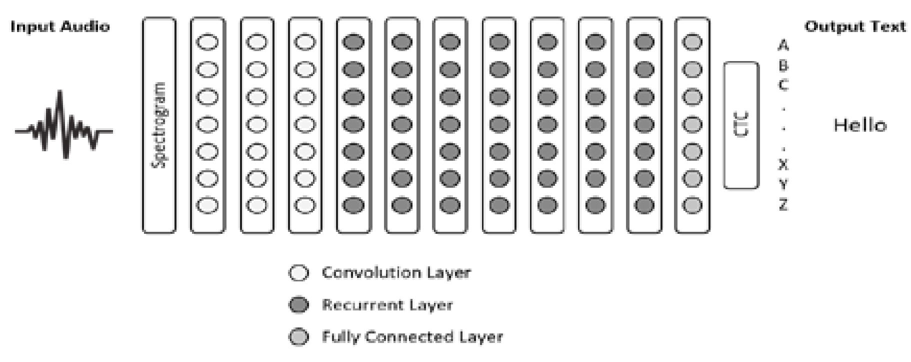


الشكل ١٣: مثال لشبكة ترجمة آلية

٦,٤ التعرف على الكلام (Speech Recognition)

تحويل الكلام المنطوق من موجات صوتية إلى نص مكتوب يستخدم الآن في العديد من التطبيقات كالمساعدات الشخصية وتحويل الكلام المسجل إلى نصوص. وقد كان الاعتماد سابقاً بشكل كبير على خوارزميات HMM، ولكن في الآونة الأخيرة تم تبني التعلم العميق بشكل أساسي، حيث تستخدمه الآن كبرى الشركات في منتجاتها للتعرف على الكلام. أحد الخوارزميات المشهورة هي DeepSpeech (Amodei، وآخرون، ٢٠١٦) الموضحة في الشكل ١٤. في البداية يتم تحويل المقطع الصوتي إلى

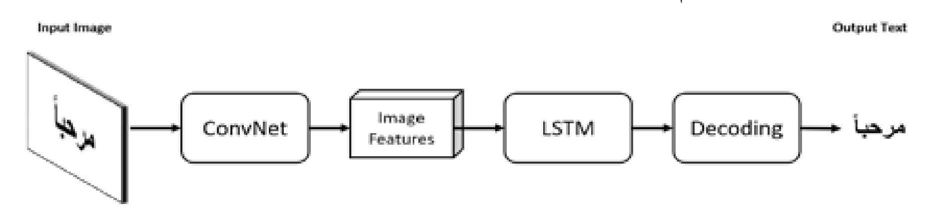
الطيفية (spectogram) -وهو تمثيل للتردد عبر الزمن- ثم استخدام عدة طبقات من الشبكات العصبية الترشيحية، متبوعة بطبقات من نوع RNN، ثم طبقة تامة الاتصال. الطبقة الأخيرة المستخدمة تدعى Connectionist Temporal Classification (CTC) (Graves، Fernández، Gomez، و Schmidhuber، ٢٠٠٦) وفيها يتم اختيار المخرجات الأعلى احتمالاً.



الشكل ١٤: شبكة Deep Speech للتعرف على الكلام

٦,٥ تحويل الصور إلى نصوص (Optical Character Recognition)

من التطبيقات المهمة تحويل النصوص المطبوعة أو المكتوبة يدوياً إلى نص قابل للتعديل على الحاسب. ولهذا المجال العديد من التطبيقات، كالفرز الآلي للطرود، قراءة أرقام الشيكات، وقراءة أرقام لوحات السيارات.



الشكل ١٥: مثال لشبكة لتحويل الصور إلى نصوص

بالرغم من الاختلاف الظاهري بين مشكلتي التعرف على الكلام والتعرف على الكتابة في الصور، فإن فكرة معمارية شبكة التعلم العميق مشابهة جداً للتي تم شرحها للتعرف على الكلام كما هو موضح في الشكل ١٥. ففي البداية يتم استخدام الشبكات العصبية الترشيحية، متبوعة بطبقات من نوع RNN، ثم طبقة تامة الاتصال، ثم خوارزمية لفك الترميز مثل CTC التي تم شرحها في خوارزمية التعرف على الكلام.

٦,٦ توليد الكلام (Speech Synthesis)

عكس تحويل الكلام إلى نصوص، الهدف من توليد الكلام هو تحويل النص المكتوب إلى مقطع صوتي منطوق. كان في السابق يتم إلصاق الفونيمات (الوحدات الصوتية) لإنشاء الكلام، ولكن في السنوات الأخيرة تم ابتكار العديد من الخوارزميات باستخدام التعلم العميق تعطي نتائج مقاربة بشكل كبير للصوت البشري. من الخوارزميات المهمة لتوليد الكلام باستخدام التعلم العميق هي خوارزمية wavenet من قوقل (Oord، وآخرون، ٢٠١٦)، وهي خوارزمية توليدية تتعلم التوزيعة المشروطة التالية:

$$p(\boldsymbol{x}) = \prod_{t} p(x_t | x_{<t}, \boldsymbol{\theta})$$

بحيث x_t هو المتغير t، و$\boldsymbol{\theta}$ مدخلات (parameters) النموذج. في هذا النموذج يتم توليد العينة الصوتية x_t بناء على ما يسبقها من العينات $x_{<t}$. مشكلة توليد الصوت بهذه الطريقة أنه يتطلب الكثير من المعالجة لأن معدل العينات في المقاطع الصوتية عالٍ جداً. لهذا كانت wavenet بطيئة جداً في البداية، ولكن تم تحسينها لاحقاً حتى وصلت إلى مستوى أداء مقبول.

٦,٧ المزيد من التطبيقات

هناك المزيد من تطبيقات معالجة اللغات التي تم استخدام التعلم العميق فيها وأعطت نتائج ممتازة. يصعب أن نحصرها في هذه المقدمة البسيطة، ولكن فيما يلي أمثلة على تطبيقات لم تذكر هنا:

- تصنيف النصوص (Text Classification)
- تلخيص النصوص (Text Summarization)
- الإجابة على الأسئلة (Question and Answering)
- التعرف على الأعلام ((Named Entity Recognition (NER))
- الكشف عن النسخ المعدل (Paraphrase Detection)
- التصحيح الإملائي (Spell Checking)
- توليد النصوص (Natural Language Generation)

المراجع

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., . . . Zhu, Z. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, (pp. 173-182).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 135-146.

Cho, K., Merrienboer, B. v., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machinetranslation: Encoder-decoder Approaches. arXiv preprint.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd International Conference on Machine learning (ICML '06), (pp. 369-376).

Hebb, D. (1949). The Organization of Behavior. New York: Wiley.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 1527-1554.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput., 1735-1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, (pp. 554-2558).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, (pp. 1097-1105).

McCulloch, W., & Walter, P. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 115–133.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, (pp. 3111-3119).

Minsky, M., & Papert, S. (1969). Perceptrons: An Introduction to Computational Geometry. MIT Press.

Oord, A. v., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., . . . Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. SSW.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. In EMNLP.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain. Psychological Review, 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations by Error Propagation. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 318-362.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, (pp. 3104-3112).

Widrow, B. (1960). An Adaptive Adaline Neuron Using Chemical Memistors. Stanford Electronics Laboratories Technical Report.

Winter, R., & Widrow, B. (1988). MADALINE RULE II: A training algorithm for neural networks. IEEE International Conference on Neural Networks, 401-408.

الفصل الثالث

الترجمة الآلية

د. عبدالله بن صالح الراجح

ملخص الفصل

نشهد حالياً تطوراً ملحوظاً في أداء أنظمة الترجمة الآلية بعد عقود من البحث والتطوير، مما ساهم في زيادة الاعتماد عليها من المستخدم العادي وكذلك المترجم المحترف. لقد ساهمت هذه الأنظمة في تسهيل الوصول للمعرفة بشتى اللغات وكذلك التواصل مع الأمم الأخرى بأقل التكاليف. وتعد أتمتة الترجمة من أصعب المشاكل في مجال الذكاء الاصطناعي حيث تتطلب معارف لغوية على عدة مستويات لمحاكاة عمل المترجم المختص. يقدم هذا الفصل نظرة عامة على مجال الترجمة الآلية وتاريخه وأهم الأبحاث المقدمة فيه خصوصاً المتعلقة بترجمة اللغة العربية. كما يستعرض منهج الترجمة الآلية الإحصائية (Statistical Machine Translation) والذي كان المهيمن على مدى عدة عقود من الزمن إلى أن تحول المجتمع البحثي حديثاً ولحقه كبريات الشركات إلى المنهج المعتمد على الشبكات العصبية (Neural Machine Translation). وبهذه النقلة النوعية دخلت الترجمة الآلية عصراً جديداً سيتم عرض أهم ملامحه. وبالرغم من النجاحات إلا أن هناك العديد من التحديات التي سنتطرق إلى أهمها في نهاية هذا الفصل.

د. عبدالله بن صالح الراجح

حاصل على درجة الدكتوراه في علوم الحاسب من جامعة ساوثامبتون في بريطانيا عام ٢٠١٥م ودرجة الماجستير في علوم الحاسب من جامعة مانشستر في بريطانيا عام ٢٠٠٩م ودرجة البكالوريوس في علوم الحاسب من جامعة الملك سعود عام ٢٠٠٦م. يعمل أستاذ بحث مساعد في المركز الوطني لتقنية الذكاء الاصطناعي والبيانات الضخمة بمدينة الملك عبدالعزيز للعلوم والتقنية. نشر العديد من الأبحاث في مجال تعليم الآلة وتطبيقاته في معالجة اللغات الطبيعية. وعمل على عدة مشاريع منها التعرف الضوئي على الكتابة العربية، وكذلك كتابة برايل وأيضاً تصنيف النصوص العربية كما أدار مشروع الترجمة الآلية من العبرية إلى العربية ويعمل حالياً على مشروع المساعد الافتراضي العربي. اهتماماته البحثية تتركز في الترجمة الآلية ومعالجة الكلام باستخدام التعلم العميق. (asrajeh@kacst.edu.sa)

١ - مقدمة

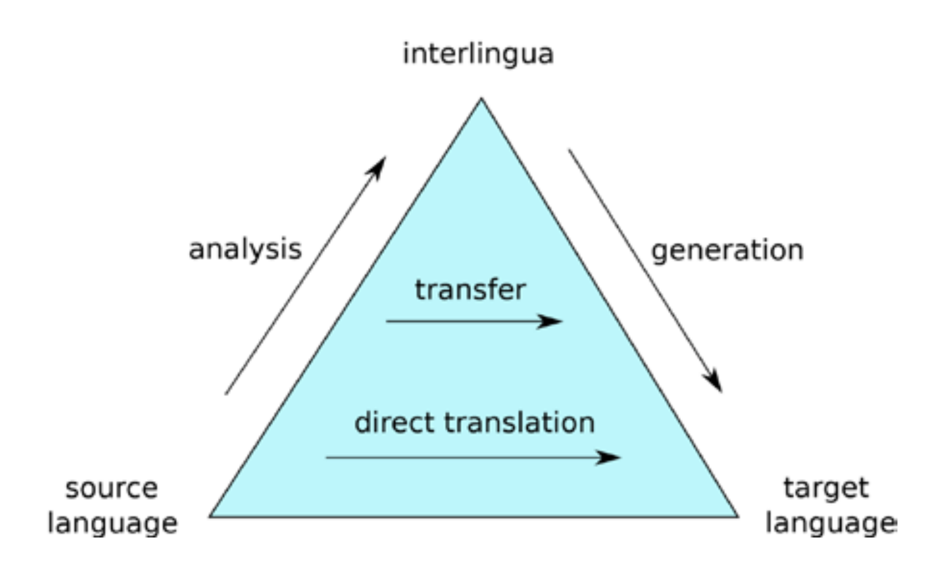
لقد ساهمت الترجمة وتساهم في نقل الثقافات والعلوم بين الشعوب وتسهيل التواصل فيما بينها. ويبذل المترجمون جهداً ووقتاً كبيرين في ذلك. حيث إن مجرد القدرة على التحدث باللغة المترجم إليها لا تكفي، بل يتطلب الأمر معارف أخرى تدرس في الجامعات والمعاهد المتخصصة. ويعد المترجم المتمكن عملة نادرة خصوصاً في مجال الترجمة الأدبية التي تتطلب فهماً أعلى للغة، وكذلك الترجمة الفورية لما تفرضه من سرعة. ولذا فإن عملية الترجمة مكلفة على جميع المستويات.

ومع بداية ظهور الحواسيب برزت مشكلة الترجمة الآلية (Machine Translation) لمحاكاة عمل المترجم وهي إحدى أقدم وأصعب المشاكل في مجال الذكاء الاصطناعي (Artificial Intelligence). وقد بُذلت الجهود للمساهمة في حلها مدفوعة برغبة أجهزة الاستخبارات مضاعفة قدراتها في جمع المعلومات عن الدول الأجنبية. وبالرغم من التاريخ الطويل إلا أن الأتمتة الكاملة للترجمة بجودة عالية لا تبدو قريبة المنال نظراً لارتباط الترجمة بقضايا لغوية وثقافية تصعب على الإنسان فكيف بالآلة. يجدر بالذكر أن هناك تطوراً ملحوظاً في جودة الترجمة الآلية المعتمدة على منهجية التعلم العميق (Deep Learning) مما ساهم في زيادة الاعتماد على أنظمة الترجمة من المستخدم العادي وكذلك المترجم المحترف.

يمكن لنا تعريف الترجمة ببساطة بأنها عملية نقل معنى النص من لغة إلى أخرى. وهذه العملية تتطلب مجموعة من المهارات بدءاً بالمعرفة الكاملة للغة الأصل (Source Language) على جميع المستويات من صرف (Morphology) ونحو (Syntax) ومعانٍ (Semantics) وتأويل (Pragmatics) ومعرفة بسياق النص المترجم (Context) وانتهاءً بمعرفة مماثلة للغة المترجم إليها (Target Language).

وهناك عدة مناهج للترجمة تتدرج في مستويات التعقيد من الترجمة المباشرة (Direct) إلى مستوى النقل (Transfer) من خلال التحليل الصرفي والنحوي وانتهاءً بمستوى تجريد المعنى عن طريق لغة عالمية مستقلة (Interlingua) ثم صياغته إلى اللغة الأخرى (Vauquois, 1968). المخطط الهرمي لفاكويس يوضح مناهج الترجمة (الشكل ١).

نسعى في هذا الفصل إلى إعطاء القارئ الغير متخصص نظرة عامة عن الترجمة الآلية بدءاً من تاريخها ثم أهم المناهج المتبعة لبناء أنظمة الترجمة وكيفية تقييم جودتها. ثم سنتحدث عن عصر جديد تعيشه الترجمة الآلية مع دخول تقنيات التعلم العميق وما واكبها من تطور في جودة الترجمة. أخيراً سنتطرق إلى أبرز التحديات التي يواجهها الباحثون في هذا المجال. وسيكون التركيز الأكبر خلال الفصل على أنظمة الترجمة من اللغة العربية وإليها.



الشكل ١: مخطط فاكويس الهرمي لمناهج الترجمة.

٢- شيء من التاريخ

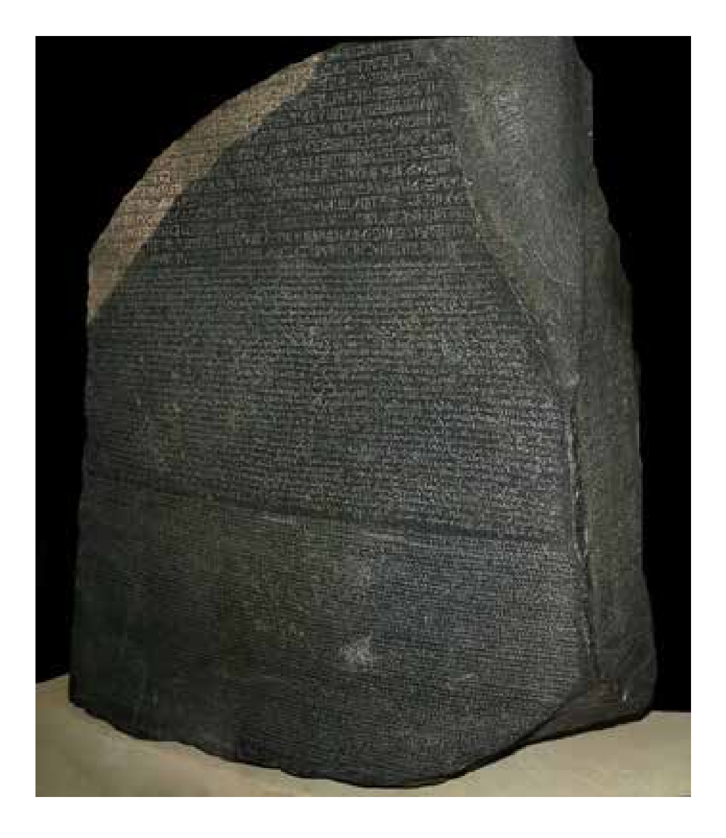
بدأ البحث في الترجمة الآلية مع ظهور الحواسيب. وكانت بريطانيا تستخدمها في الحرب العالمية الثانية لفك شفرة إنجما الألمانية (Enigma machine) الأمر الذي يعد شبيهاً بعمل الترجمة الآلية. كتب وارن ويفر، أحد الرواد في المجال، في عام ١٩٤٧م رسالة إلى نوربرت وينر يقول فيها: «عندما أنظر إلى مقال بالروسية أقول هذا مكتوب بالإنجليزية لكنه مشفر. سأقوم الآن بفك تشفيره» (Weaver, 1947).

عدد من الدول في ذلك الوقت بالذات الولايات المتحدة الأمريكية لديها رغبة لتطوير أنظمة ترجمة لأغراض أمنية وكان هناك تفاؤل كبير بحل مشكلة الترجمة الآلية في غضون سنوات. في عام ١٩٥٤م قامت جامعة جورجتاون مع شركة آي بي إم بتجربة بناء نظام ترجمة من اللغة الروسية إلى الإنجليزية اعتماداً على قاموس محدود وست قواعد لغوية فقط (Slocum, 1985). لاقت هذه التجربة أصداء واسعة جذبت اهتمام ودعم المؤسسات الحكومية الأمريكية، إلا أن التقدم بعد ذلك أصبح بطيئاً ليتم تشكيل لجنة حكومية (ALPAC) بعد عقد من الزمان لتقييم أبحاث الترجمة الآلية. في عام ١٩٦٦م خلصت اللجنة في تقريرها إلى أن قدرات الترجمة الآلية مبالغ فيها وأن تكاليف المترجمين أقل من تحرير مخرجات أنظمة الترجمة ونتيجة لذلك توقف تمويل أبحاث الترجمة الآلية في أغلبه (Philipp, 2010).

بعد عدة سنوات عادت الأبحاث لتركز على التمثيل المجرد للمعنى (meaning-oriented) بشكل مستقل عن اللغة المحددة. وبالرغم من جاذبية الفكرة إلا أن صعوبة تنفيذها حال دون إحراز تقدم فيها وعدت من المشاكل الكبرى في الذكاء الاصطناعي. على العكس من ذلك كان هناك تقدم في أنظمة الترجمة المعتمدة على قواعد اللغة (rule-based) المبنية من قبل مختصين في اللغة (linguistics). كانت هذه الأنظمة فعالة لأن اللغة في مجملها ثابته (static) إلا أن بناءها مكلف مادياً ويستغرق وقتاً لحصر القواعد من الخبراء لكل لغة جديدة. الأمر الآخر إضافة قواعد لغوية قد يخلق تعارضات مع القواعد السابقة و يتطلب حلها وقتاً طويلاً. من أشهر الأنظمة التجارية في تلك الفترة (Systran) و (Logos).

٣- حجر رشيد

عادةً ما يُرمز بحجر رشيد (Rosetta Stone) للمنهج الحديث في الترجمة الآلية المعتمد على نصوص مترجمة سابقة (data-driven approach). أُكتشف الحجر في مصر عام ١٧٩٩م جنوب الجيزة، منقوش عليه مرسوم ملكي بالمصرية واليونانية القديمة يعود لعام ١٩٦ قبل الميلاد في عهد الملك بطليموس الخامس (الشكل ٢). كان اكتشافه مفتاحاً لفك شفرة الهيروغليفية المصرية على معابد ومقابر الفراعنة.



الشكل ٢: حجر رشيد منقوش عليه مرسوم ملكي بالمصرية القديمة في الأعلى والوسط وباليونانية القديمة في الأسفل معروض في المتحف البريطاني (Wikipedia, © Hans Hillewaert).

تمكن الباحثون بعد ٢٠ عاماً من فك شفرة اللغة المصرية القديمة عن طريق اللغة اليونانية القديمة التي كانت معروفة من خلال المقارنة بين الثلاث نسخ لنص المرسوم الملكي. وهنا تكمن رمزية حجر رشيد للباحثين في الترجمة الآلية حيث أنه يمكن تعلم ترجمة اللغات من خلال توفر نصوص مترجمة متقابلة وكلما زادت النصوص سهل تعلم الترجمة.

٤- الترجمة الآلية الإحصائية

كما ذكرنا سابقاً فإن المنهجية الحديثة للترجمة الآلية تعتمد على تعلم الترجمة من خلال نصوص مترجمة سابقاً. هذه المنهجية بدأت تكتسب زخماً في نهاية الثمانينيات الميلادية حتى وقتنا الحاضر. فبدلاً من الاستعانة بخبراء اللغة لكتابة قواعد الترجمة كما في الأنظمة

القائمة على القواعد (rule-based) يمكن استخراج قواعد احتمالية (probabilistic rules) من النصوص من خلال الإحصاء. ففي عام ١٩٩٣م نشر باحثون من شركة آي بي إم ورقة علمية مشهورة بعنوان «رياضيات الترجمة الآلية الإحصائية» تصف خمسة نماذج لبناء نظام ترجمة من اللغة الفرنسية إلى الإنجليزية عرفت لاحقاً بنماذج آي بي إم (Brown et al., 1993) اعتماداً على نصوص وقائع البرلمان الكندي المدونة باللغتين. وبعد سنوات قام باحثون ببرمجتها وجعلها مفتوحة المصدر أثناء ورشة صيفية في جامعة جونز هوبكنز (Al-Onaizan et al.,1999).

يقترح (Brown et al., 1993) أن أفضل ترجمة لجملة فرنسية معطاة f إلى جملة إنجليزية e هي التي تزيد من قيمة الاحتمال المشروط كالتالي:

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$$

وحيث إن هناك عدداً لا محدوداً من الجمل الإنجليزية، فإنه من الصعب بناء نموذج واحد يميز بينها. لذلك يتم تقسيم المشكلة إلى أجزاء أسهل باستخدام قانون بيز (Bayes rule) لتصبح كالتالي:

$$\text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}}\, \frac{p(\mathbf{e})p(\mathbf{f}|\mathbf{e})}{p(\mathbf{f})}$$

حيث إن النموذج $p(\mathbf{f}|\mathbf{e})$ يعطي احتمالية أن الجملة تحمل المعنى الصحيح (translation model) والنموذج $p(\mathbf{e})$ يعطي احتمالية أن الجملة سليمة لغوياً (language model). فبدلاً من البحث عن ترجمة صحيحة وخالية من الأخطاء اللغوية في وقت واحد، يتم التركيز على الجمل السليمة لغوياً ونتجاهل البقية لضعف احتمال وقوعها. وهذه الطريقة مشهورة في مجال الاتصالات وتسمى (noisy-channel model) التي تفترض أن شخصاً يتلقى رسائل من صديقه، بعضها يصل مشوهاً، ولاستعادة الرسائل الأصلية يتم البحث عن أكثر الرسائل المحتملة من صديقه، والتي يمكن أن تُشوه بهذه الطريقة من خلال الخبرة السابقة.

ويمكن تقدير سلامة الجملة لغوياً من خلال حساب احتمالية وقوعها بعد تجزئتها إلى كلمات، وحساب احتمالية كل كلمة مشروطة بما سبقها باستخدام قاعدة السلسلة (chain rule) كالتالي:

$$p(\mathbf{e}) = p(e_1.e_2.\cdots.e_I) = p(e_1)p(e_2|e_1)\cdots p(e_I|e_1.\cdots.e_{I-1})$$

إلا أن هناك عدداً لا محدوداً من السياقات التي يمكن أن تقع فيها كل كلمة، لذلك عادةً ما يُحصر السياق بعدد محدود من الكلمات (n)، عادةً من ثلاث إلى خمس كلمات، ويسمى هذا النموذج (n-gram model) حيث إن احتمالية كل كلمة تحسب كما يلي:

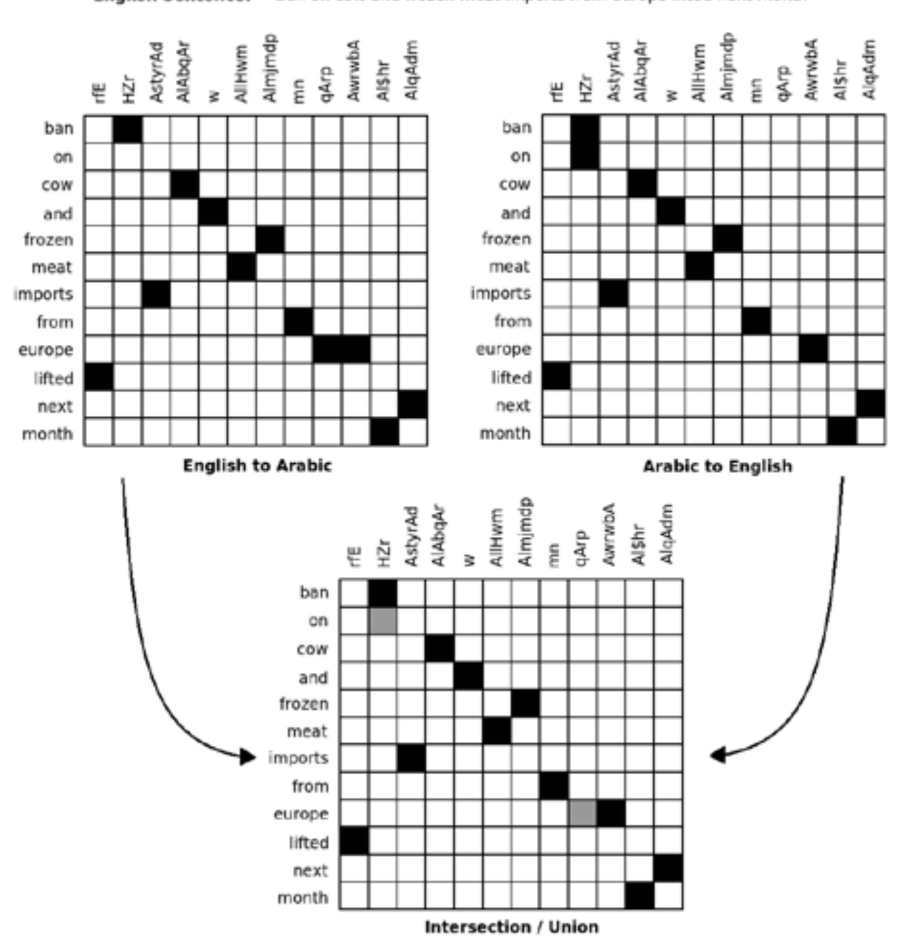
$$p(e_i|e_1.\cdots.e_{I-1}) \simeq p(e_i|e_{i-n}.\cdots.e_{i-1})$$

ويمكن تقدير هذه الاحتمالات من نصوص كثيرة (ملايين الجمل) من خلال احصاء كل كلمة والسياقات التي وردت فيها مقسوماً على بقية الكلمات التي وردت في نفس تلك السياقات كما يلي:

$$p(e_i|e_{i-n}.\cdots.e_{i-1}) = \frac{\text{count}(e_{i-n}.\cdots.e_{i-1}.e_i)}{\text{count}(e_{i-n}.\cdots.e_{i-1}.e)}$$

أما تقدير صحة نقل الجملة للمعنى فيتم ببساطة عن طريق الإحصاءات المعجمية (lexical statistics)، والتي تقدر من نصوص كثيرة متقابلة من اللغتين وباستخدام نماذج آي بي إم (IBM models) تقدر الاحتمالات (Brown et al., 1993). فبدلاً من الاعتماد على قواميس ثابتة يتم احتساب احتمالية ترجمة أي كلمة إلى اللغة المقابلة $p(f|e)$ وكلما كانت الترجمة صحيحة تكون أقرب إلى واحد والخاطئة أقرب إلى صفر.

إن هذه الطريقة تُبسط عملية الترجمة بافتراض أنها تقع على مستوى الكلمات (ward-based)، إلا أنها لا تأخذ السياق في الحسبان، فبعض الكلمات تُترجم معاً مما يجعل منهج ترجمة العبارات (phrase-based) أفضل من ترجمة الكلمات (Koehn, 2010). وعادةً ما تُستخرج العبارات من معرفة محاذاة الكلمات (word alignments) عن طريق نماذج آي بي إم كما يبين الشكل ٣. تجدر الإشارة إلى أن هناك طريقة أخرى تستخرج عبارات هرمية (hierarchical phrases) لا يتسع الفصل لشرحها (Chiang, 2007).



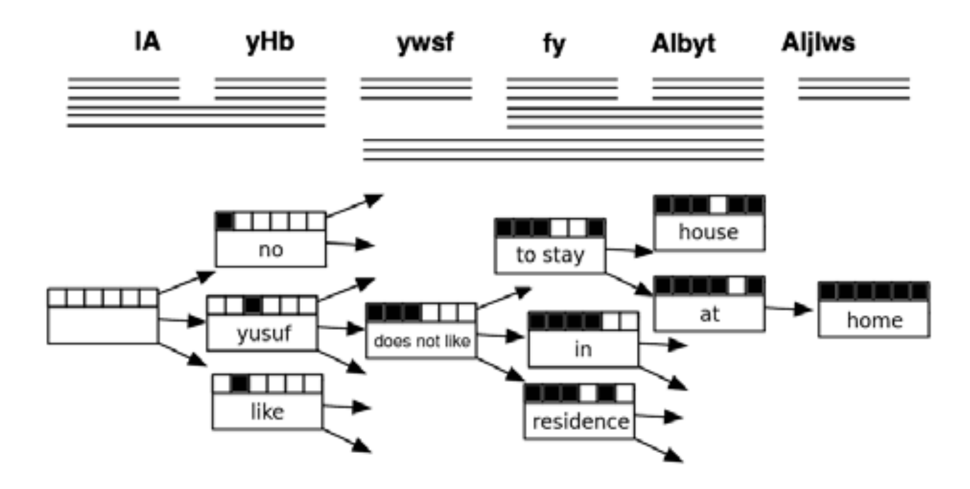
الشكل ٣: محاذات الكلمات لجملة عربية مع إنجليزية من خلالها يتم استخراج ترجمة العبارات.

بعد استخراج العبارات من جميع الجمل المتقابلة يتم بسهولة حساب احتمال ترجمة كل عبارة من خلال التكرار النسبي (relative frequency) وتوضع في جدول ضخم (translation table). عادة يتم اعتبار ترجمات محدودة لكل عبارة (عشرين مثلاً) أثناء البحث (decoding) عن الترجمة الصحيحة (الشكل ٤). لاحظ أن إعادة ترتيب العبارات لتكون جملة سليمة لغوياً من مهام نموذج اللغة كما تم ذكره سابقاً.



الشكل ٤: توضيح لخيارات البحث أثناء ترجمة جملة عربية إلى الإنجليزية.

إن مهمة البحث عن أفضل العبارات والكلمات لترجمة جملة ما ليست سهلة. ومن أشهر خوارزميات البحث الفعالة ما يعرف بالبحث الشعاعي (beam search) الذي يستكشف أفضل الخيارات، لكنه لا يضمن الحل الأفضل. ويبدأ ببناء ترجمات جزئية تعرف بفرضيات (hypotheses) ثم يوسع كل فرضية بشكل محدود حتى يصل إلى نهاية الجملة (الشكل ٥) وأفضل فرضية هي التي تحقق أعلى احتمالية اعتماداً نموذجي اللغة والترجمة.

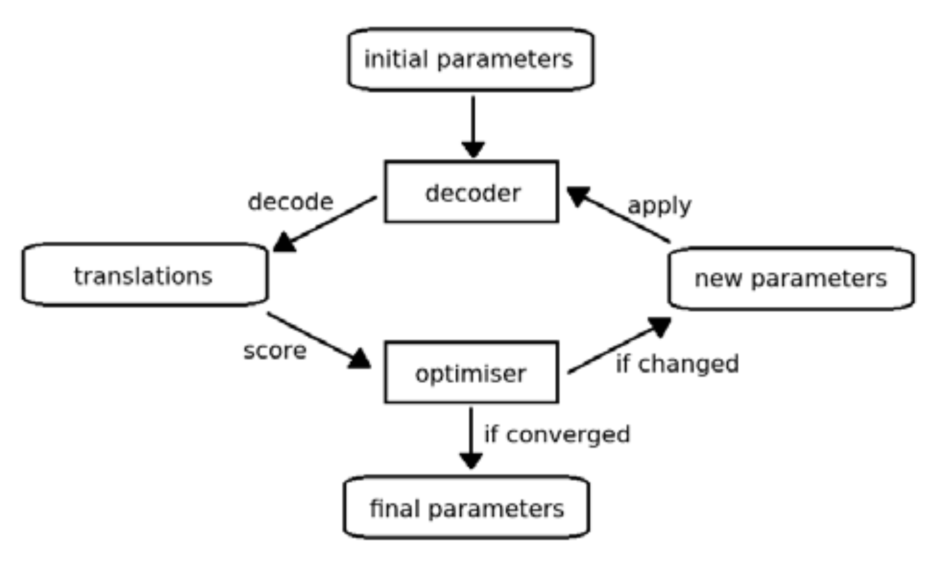


الشكل ٥: بحث شعاعي (beam search) عن أفضل ترجمة لجملة عربية.

وبالرغم من أن أنظمة الترجمة تقوم على نموذجي اللغة والترجمة إلا أن الباحثين قاموا بإضافة العديد من الأجزاء التي تحسن الترجمة من خلال إطار يسمى (log-linear framework) يُمَكّن من إضافة أجزاء أخرى لنظام الترجمة $p_i(\mathbf{f.e})$ وإعطائها وزناً محدداً λ_i يعكس أهميتها كالتالي:

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}} \sum_{i \in \{\text{t,lm,lex,d,w}\}}^{n} \lambda_i * \log\ p_i(\mathbf{f.e})$$

وعادة ما تحتوي أنظمة الترجمة الإحصائية كنظام موسز (Moses) مفتوح المصدر (koehn, 2007) على خمسة أجزاء هي (translation model) و(language model) و(lexical model) و(reordering model) و(word penalty). وكل جزء يمكن إعطاؤه وزناً اعتباطياً، إلا أن تحديدها يكون عادة من خلال اختبار النظام على مجموعة من الجمل المترجمة مُسبقاً بعدة أوزان، ومن ثم اختيار الأفضل (discriminative training) وهناك العديد من الخوارزميات التي تقوم بذلك أشهرها (MERT) الموضحة في الشكل ٦ (Och, 2003).



الشكل ٦: طريقة ضبط الأوزان لأجزاء نظام الترجمة.

هذا وقد كان لمدينة الملك عبدالعزيز للعلوم والتقنية جهود في بناء أنظمة ترجمة لخدمة المملكة. كان بدايتها التعاون مع شركة آي بي إم الرائدة في هذا المجال عام 2009م. وخلال هذا التعاون تم بناء نظام ترجمة من العبرية إلى العربية والفارسية إلى العربية بجودة منافسة للأنظمة التجارية نظراً لاعتمادهما على نصوص مترجمة بجودة عالية تجاوزت ٥ مليون كلمة لكل لغة. الشكل ٧ يوضح واجهة النظام على الشبكة.



الشكل ٧: واجهة نظام ترجمة من العبرية والفارسية إلى العربية (translate.kacst.edu.sa).

٥- تقييم جودة الترجمة

يعد تقييم أداء أنظمة الترجمة أمراً صعباً وذلك لتعدد الترجمات من شخص لآخر (subjective). المثال التالي دائماً ما يذكر لتوضيح المشكلة وهو مأخوذ من مجموعة تقييم نيست (NIST) لعام ٢٠٠١م. لاحظ أن هناك عشر ترجمات مقبولة لهذه الجمل باللغة الصينية.

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli o-cials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security o-cials.

هناك معياران لتقييم الترجمة هما مدى الدقة في نقل المعنى (adequacy) ومدى سلاسة الترجمة (fluency). وقد تم اقتراح العديد من الأدوات لقياس دقة المعنى والسلاسة يمكن تصنيفها إلى مجموعتين (manual metrics) و(automatic metrics). وتعد المجموعة الثانية عملية أكثر وأقل كلفة نظراً لغياب العنصر البشري فيها وثبات النتائج عند إعادة القياس (consistent). حيث إنها تعتمد على ترجمات احترافية (references) سابقة للنصوص المراد قياس أداء النظام فيها. ومن أبسط أدوات القياس (precision) و (recall) والتي يمكن حسابها كالتالي:

$$\text{precision} = \frac{\text{correct words}}{\text{translation length}}$$

$$\text{recall} = \frac{\text{correct words}}{\text{reference length}}$$

إلا أن أكبر نقاط ضعفها تجاهل ترتيب الكلمات والذي يعد أساساً في سلاسة الترجمة. ويمكن معالجة ذلك من خلال أداة (WER) والتي تقيس الحد الأدنى من الخطوات اللازمة لتحرير ترجمة النظام لتصبح مثل الترجمة الاحترافية كالتالي:

$$\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference length}}$$

ومن أشهر أدوات القياس حالياً بلو (BLEU) من مركز أبحاث واتسون في آي بي إم والتي تستخدم في أغلب أبحاث الترجمة الآلية رغم أنها من أوائل ما تم اقتراحه (Papineni et al., 2002). وتقوم على قياس دقة الترجمة على مستوى العبارات (n-grams) كما يلي:

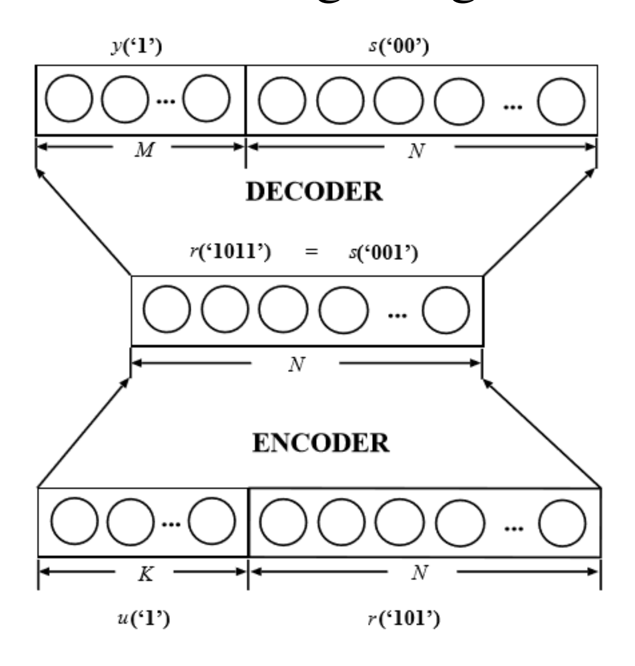
$$\text{BLEU} = \text{BP} * \exp \sum_{i=1}^{n} \lambda_i * \log(\text{precision}_i)$$

$$\text{BP} = \min(1, \frac{\text{translation length}}{\text{reference length}})$$

وتعرف (BP) على أنها عقوبة الإيجاز فكلما كان طول الترجمة أقصر من الترجمة الاحترافية نقصت نقاط بلو والتي تصل إلى ١٠٠ نقطة عند مطابقة ترجمة النظام للترجمات الاحترافية. وتحقيق كل نقطة ليس بالأمر السهل، حيث إن أفضل أنظمة الترجمة تصل إلى ٦٠ نقطة (Junczys-Dowmunt et al., 2016).

٦- عصر جديد

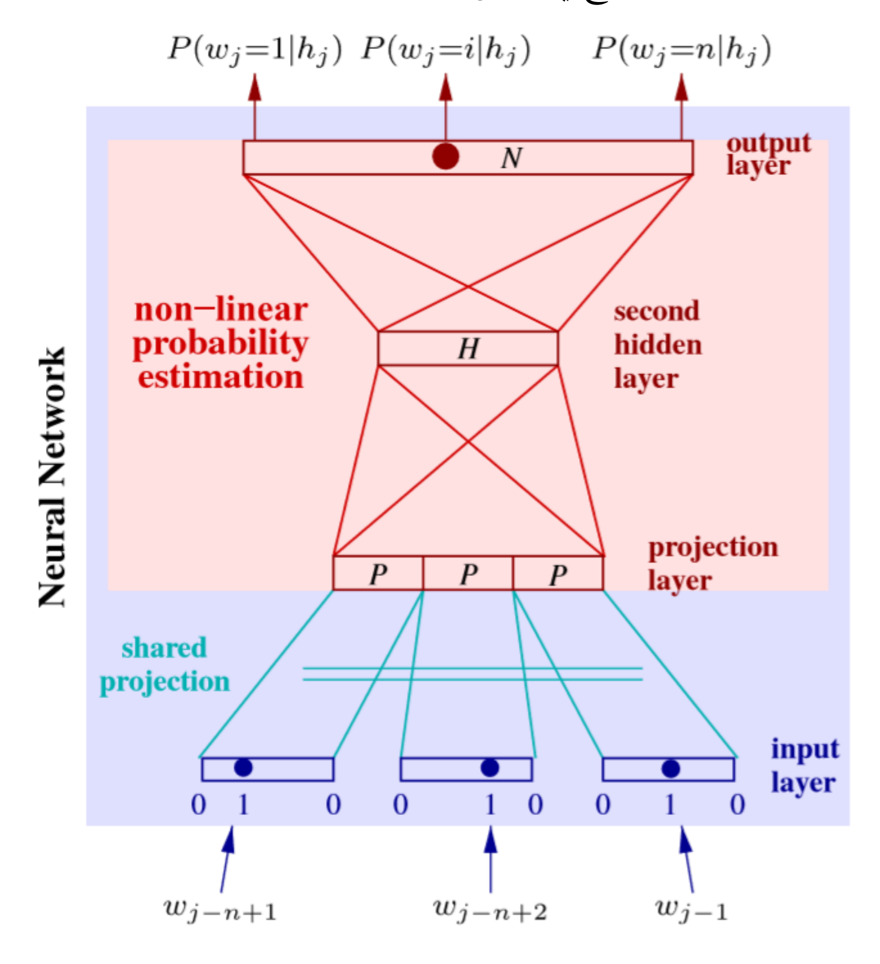
خلال السنوات القليلة الماضية حدث تغير جذري في أبحاث الترجمة الآلية (paradigm shift) من المنهج الإحصائي إلى ما يعرف بالترجمة الآلية العصبية (neural machine translation) المعتمدة على الشبكات العصبية العميقة (deep neural networks) في ترجمة كامل الجملة باستخدام نموذج واحد متكامل (end-to-end system). إن استخدام الشبكات العصبية ليس بالأمر الجديد، فقد تم اقتراح نماذج مشابهة لما هو معمول به الآن قبل أكثر من عقدين من الزمن (Forcada, 1997) كما في الشكل ٨. إلا أن تعقيدها تطلب حواسيب قوية لتدريبها على بيانات كافية وهو ما لم يكن متوفراً. لذلك كانت نتائج تلك النماذج ضعيفة مما أدى إلى هجران تلك الأفكار.



الشكل٨: بنية لنظام ترجمة من مرحلتين: تشفير ثم فك التشفير (Forcada, 1997)

مع مرور السنوات زادت سرعة الحواسيب وبدأت تستبدل النماذج العصبية أجزاء من أنظمة الترجمة، كورقة هولغر شونك (Schwenk, 2007) عن نمذجة اللغة في فضاءات مستمرة اعتماداً على فكرة (Bengio and Ducharme, 2001) والتي يقوم جوهرها على تمثيل الكلمات بمتجهات (vectors) ذات معنى دلالي بدلاً من أرقام

اعتباطية يمكن تعلمها من نصوص كثيرة عُرفت لاحقاً بتضمين الكلمات (word embedding) كما هو موضح في الشكل ٩.



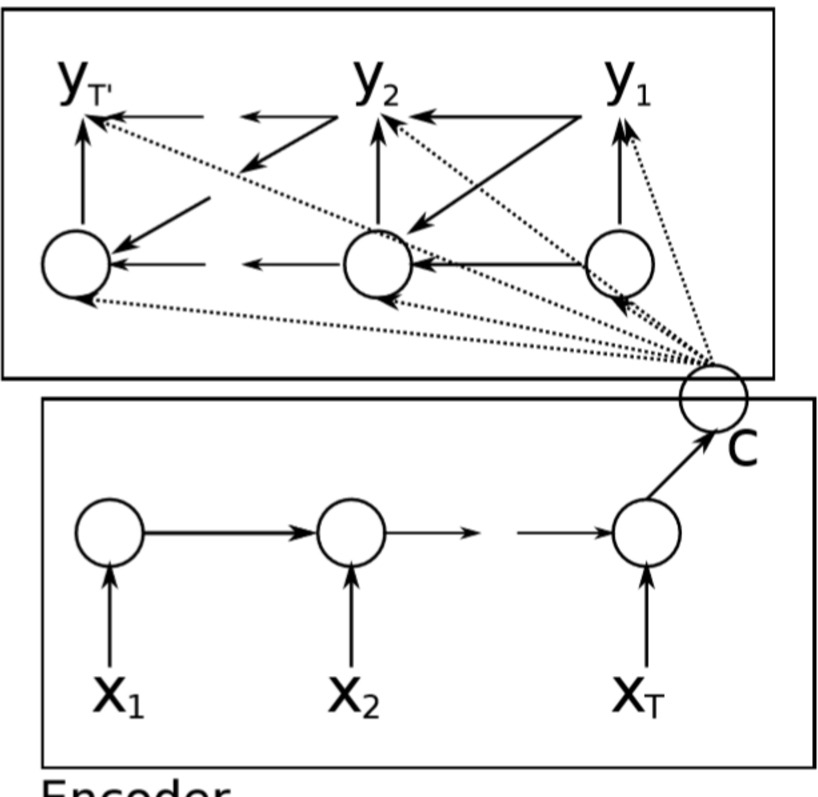
الشكل ٩: بنية شبكة عصبية لنمذجة اللغة من خلال تمثيل سياق الكلمة في متجه واحد (projection) ومن ثم حساب احتمالية الكلمة بناءً عليه (Schwenk, 2007)

وقد أظهرت التجارب تحسناً كبيراً في الترجمة، إلا أن تبنيها كان محدوداً نظراً للكلفة الحوسبية القائمة بشكل رئيس على حساب المصفوفات. هذا وقد ظهرت تجارب لتدريب هذه الشبكات العصبية على وحدات معالجة الرسومات (GPUs) السريعة في معالجة المصفوفات، إلا أن عدم توفرها لكثير من الباحثين حال دون انتشارها.

وقد كان لورقة (Devlin et al., 2014) أثر على المجتمع البحثي للنتائج القوية التي عرضتها ومنحت جائزة أفضل ورقة في مؤتمر (ACL). فقد أظهرت قدرة نماذج اللغة المبنية على الشبكات العصبية على تحسين أفضل أنظمة الترجمة من اللغة العربية والصينية إلى الإنجليزية. ففي حملة (NIST OpenMT) لتقييم أنظمة الترجمة عام ٢٠١٢م حقق المركز الأول في الترجمة من العربية إلى الإنجليزية ٤٩,٥ نقطة بلو (BLEU) واستطاعت الورقة تخطي ذلك بأكثر من ٣ نقاط محققةً ٥٢,٨ نقطة.

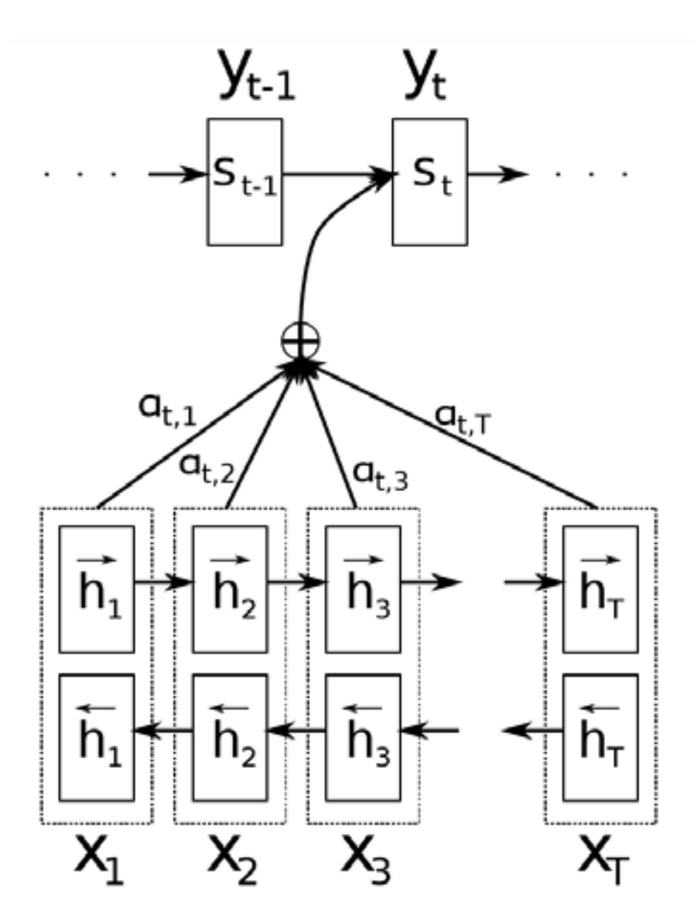
ويمكن اعتبار ورقة (Kalchbrenner and Blunsom, 2013) حجر الأساس لأنظمة الترجمة الآلية العصبية من خلال طرح نموذج متكامل للترجمة (end-to-end encoder-decoder). وقد تم استخدام شبكات عصبية التفافية (convolutional neural networks) لتشفير الجملة المراد ترجمتها (encoding) ومن ثم فكها (decoding) لتوليد الترجمة من خلال شبكات عصبية متكررة (recurrent neural networks).

وبالرغم من نجاحات الشبكات العصبية إلا أنها لم تستطع التفوق على المنهج التقليدي في ترجمة الجمل الطويلة. وقد طُرحت العديد من الحلول أبرزها استخدام (LSTM or GRU units) وهي عبارة عن وحدات عصبية قادرة على التذكر (Sutskever et al., 2014; Cho et al., 2014). إلا أن النموذج في ذلك الوقت كان قائماً على تشفير الجملة المراد ترجمتها إلى متجه ذي حجم ثابت (الشكل ١٠)، سواءً طالت الجملة أم قصرت، وهو ما عُد عقبة أمام ترجمة الجمل الطويلة.



الشكل ١٠: شبكة عصبية متكررة لشفير الجملة في متجه ثابت الحجم ثم ترجمتها (Cho et al., 2014)

هذا وقد قام باحثون بابتكار آلية الانتباه الفعالة (attention mechanism) الموضحة في الشكل ١١ والتي تخطت عقبة ترجمة الجمل الطويلة (Bahdanau et al., 2015). وخلال سنتين تحول المجتمع البحثي للمنهج الجديد القائم على الشبكات العصبية. ففي عام ٢٠١٥م كان هناك نظام واحد فقط عصبي صرف (pure neural) مقدم للتقييم في مؤتمر الترجمة الآلية المعروف (WMT)، وفي عام ٢٠١٧م تحولت أغلب الأنظمة المقدمة في المؤتمر إلى الشبكات العصبية (Koehn, 2017).



الشكل ١١: شبكة عصبية للترجمة بآلية الانتباه (Bahdanau et al., 2015)

لقد كانت آلية الانتباه فعالة لدرجة أن فريقاً بحثياً من شركة قوقل نشر بحثاً يصف نموذجاً أسماه (Transformer) معتمداً عليها فقط دون الحاجة إلى شبكات عصبية متكررة (RNN) أو التفافية (CNN) مما سمح بتدريب النموذج بشكل متوازٍ (parallelization) وبوقت أقل بكثر من السابق (Vaswani et al., 2017).

ويعتبر هذا المجال البحثي نشطاً جداً، ولا يسعنا في هذا الفصل تغطيته وإنما تم ذكر أهم الأبحاث فيه. وتجدر الإشارة إلى أن مدينة الملك عبدالعزيز للعلوم والتقنية عملت مؤخراً على تجارب مكثفة لبناء أنظمة ترجمة عصبية من اللغة العربية إلى اللغة الإنجليزية والعكس كذلك، حيث إن متوسط جودة هذه الأنظمة قارب ٦٠ (Alrajeh, 2018)

نقطة بلو (BLEU). ومما يسهل على الباحثين والمطورين الاستفادة والمساهمة في هذا المجال وجود كثير من الأنظمة مفتوحة المصدر. أهم تلك الأنظمة وروابط الوصول لها كالتالي (Koehn, 2017):

Nematus (based on Tensorflow): https: / /github.com / EdinburghNLP /nematus

Marian (a C++ re-implementation of Nematus): https: / /marian-nmt. github.io /

OpenNMT (based on Torch /pyTorch): http: / /opennmt.net /

xnmt (based on DyNet): https: / /github.com /neulab /xnmt

Sockeye (based on MXNet): https: / /github.com /awslabs /sockeye

T2T (based on Tensorflow): https: / /github.com /tensorflow / tensor2tensor

٧- أبرز التحديات

رغم قدم مشكلة الترجمة الآلية والقفزات في سبيل حلها إلا أنه ما زال هناك الكثير من التحديات. وسنتطرق إلى ثلاثة تحديات تواجه المنهج الحديث (neural approach) في الترجمة (Koehn, 2017).

التحدي الأول ضعف جودة الترجمة عند عدم تطابق المجال بين النظام والنصوص المراد ترجمتها (domain mismatch). من المشاكل المعروفة أن العبارات تختلف ترجمتها من مجال لآخر فترجمة الأخبار ليست كترجمة المقالات العلمية لذلك من المهم تدريب النظام على نصوص من نفس المجال. إلا أنه كثيراً ما تتوافر النصوص خارج المجال المستهدف فيتم تدريب النظام عليها ثم تكييفه على المجال المحدد باستخدام نصوص قليلة (domain adaptation). وقد أظهر التجارب أن الأنظمة الإحصائية التقليدية تعطي نتائج جيدة خارج المجال الذي تدربت عليه بعكس الأنظمة العصبية.

التحدي الثاني الحاجة لنصوص كثيرة لتدريب النظام قبل رؤية أي تحسن (amount of training data). فرغم أن أداء الأنظمة العصبية تخطى الأنظمة الإحصائية إلا

أن ذلك مشروط بتوفر نصوص كثيرة للتدريب تتجاوز العشرة ملايين كلمة. لذلك مازالت الأنظمة العصبية تواجه تحدياً في ترجمة اللغات قليلة المصادر (low–resource languages).

التحدي الثالث حساسية النظام لنصوص التدريب التي ترجمتها غير دقيقة أو غير سليمة لغوياً (noisy data). إن الحصول على بيانات تدريب عالية الجودة مكلف للغاية لذلك أحيانا يتم الاعتماد على نصوص فيها ترجمات معيبة. ومما هو معروف عن الأنظمة الإحصائية أنها صلبة تجاه البيانات المشوشة، ففي إحدى التجارب تم تشويش نصف بيانات التدريب ومع ذلك حافظ النظام على أدائه، وما فقده أقل من نقطة بلو (BLEU) واحدة بخلاف الأنظمة العصبية التي تعتبر حساسة للتشويش.

٨- خاتمة

قدمنا في هذا الفصل نبذة مختصرة عن تاريخ الترجمة الآلية والذي بدأ مع نشوء علم الحاسب. ثم تطرقنا إلى مناهج الترجمة الآلية والتي تتدرج في مستوى معالجتها للغة بدءاً من الترجمة المباشرة إلى الترجمة التجريدية. كانت الترجمة الآلية الإحصائية أهم المناهج المهيمنة حتى وقت قريب إلى أن دخلت تقنيات التعلم العميق وأحدثت نقلة في هذا المجال دخلت معها الترجمة الآلية عصراً جديداً لا نزال نعيش أحداثه.

على مدى عدة عقود تطورت الترجمة الآلية حتى أصبحت تقنية يستخدمها الجميع ويعتمد عليها المترجمون في تسهيل عملهم. وكثير من الشركات كقوقل ومايكروسوفت تعرض خدمات الترجمة بأسعار متدنية أو مجانية لأشهر اللغات مما أتاح فرصة التواصل والاطلاع على ما عند الأمم الأخرى.

وقد تم التطرق إلى أهم الأبحاث، إلا أن هذا المجال لا زال نشطاً بحثياً، والكثير من التجارب تنشر سنوياً على عدد من اللغات كالأوربية والصينية والعربية. وما زال هناك فرص لتحسين أداء الترجمة الآلية لتجاوز التحديات الكثيرة التي تطرقنا إلى بعضها.

المراجع

Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John La-erty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. (1999). Statistical machine translation. Technical report, Johns Hopkins University, Summer Workshop.

Alrajeh, Abdullah. (2018). A Recipe for Arabic-English Neural Machine Translation. In Computing Research Repository, arXiv: 808.06116.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (ICLR).

Bengio, Yoshua, Réjean Ducharme. (2001). A neural probabilistic language model. In: Proceedings of Advances in Neural Information Processing Systems, vol. 13, 932-938.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263-311.

Chiang, David. (2007). Hierarchical phrase-based translation. Computational Linguistics, 33(2).

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724– 1734.

Devlin, Jacob, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. (2014). Fast and robust neural network joint models for statistical machine translation.

In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1370–1380.

Forcada, Mikel and Ramón Ñeco. (1997). Recursive hetero-associative memories for translation. In Biological and Artificial Computation: From Neuroscience to Technology, Springer, pages 453–462.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. In Proceedings of the 13th International Workshop on Spoken Language Translation.

Kalchbrenner, Nal and Phil Blunsom. (2013). Recurrent continuous translation models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1700–1709.

Koehn, Philipp. (2010). Statistical Machine Translation. Cambridge University Press.

Koehn, Philipp. (2017). Neural Machine Translation. In Computing Research Repository, arXiv: 1709.07809.

Koehn, Philipp and Rebecca Knowles. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondˇrej Bojar, Alexandra Constantin, and Evan Herbst. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180.

Och, Franz Josef. (2003). Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160–167.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318.

Schwenk, Holger. (2007). Continuous space language models. Computer Speech and Language, 3 (21), 492–518.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, pages 3104–3112.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez and, Lukasz Kaiser and Illia Polosukhin. (2017). Attention is All you Need. In Advances in Neural Information Processing Systems 30, 5998-6008.

Vauquois, Bernard. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In Proceedings of IFIP Congress, 1114-1122.

Weaver, Warren. (1947). Letter to Norbert Wiener.

الفصل الرابع

نمذجة الكلمة العربية

خوارزميات الذكاء الاصطناعي في تحليل الكلمة العربية لغويًا وتوزيعيًا

د. عبدالرحمن بن محمد العصيمي

ملخص الفصل

تمثل الكلمة ركيزة مهمة في فهم واستيعاب الخطاب المكتوب. فلا عجب أن نجد أبحاثًا كثيرةً تصبُّ في تحليل الجوانب المختلفة للكلمة أو تحاول تمثيل الكلمة اللغوية بشكل يفهمه الحاسب الآلي. يهدف هذا الفصل إلى بناء مقدمة لغير المتخصص لفهم أحدث الخوارزميات المستخدمة في بناء النماذج الحاسوبية للكلمة العربية الفصيحة المكتوبة. كما يحاول تفسير أسباب الصعوبات التي تكتنف نمذجة الكلمة العربية تحديدًا، بدءًا بنظامها الصرفي الغير خطي ومرورًا بغناها الصرفي وانتهاءً بمستويات الغموض العالية في النص العربي. كما يقدِّم نمطين مشهورين لتحليل الكلمة: اللغوي والتوزيعي، ويقارن بينهما، وذلك عبر مقدمة لكل نمط وتحليل الخوارزميات المستخدمة وأشهر الأدوات المتاحة. وفي الختام، نسلط الضوء مرة أخرى على قصور بعض الخوارزميات عند تحليل ونمذجة اللغة العربية، وسبل ووسائل مقترحة لمعالجة أوجه القصور.

د. عبدالرحمن بن محمد العصيمي

أستاذ مساعد في كلية الحاسب بجامعة الإمام محمد بن سعود الإسلامية، ومهتم وباحث في مجال معالجة اللغة العربية حاسوبيًا، وشغوف بالبرمجة وتطوير الأنظمة. نشر مجموعة من الأوراق في مجلات ومؤتمرات علمية، وقدم محاضرات في العديد من الملتقيات. نشر أدوات حاسوبية تُعنى بمعالجة اللغة العربية آليًا برخصة متاحة للكل.

١ – مقدمة

تطورت خوارزميات الحاسب الآلي تطورًا كبيرًا في آخرين عقدين من الزمن. فلم يعد الحاسب فقط قادرًا على تنفيذ سلسلة من العمليات التي يمليها له المبرمج، بل أصبح ذكيًا وقادرًا على اتخاذ القرار من تلقاء نفسه. ولاتخاذ القرار بشكل صحيح، لابد من طرق ووسائل لتقييم المعطيات وذلك من أجل اتخاذ أفضل القرارات.

يمكننا تعريف تعلم الآلة كالتالي: «الاستمرار في تطوير مهمة ما (م) بناء على خبرة ما (خ) باعتبار وحدة تقييم أداء معينة (ق)» (Mitchell, 1997) . فمثلا إذا كانت المهمة (م) هي التعرف على جنس الإنسان في صورة ما، والخبرة (خ) التي اكتسبها الحاسب عبر إعطائه مجموعتين من الصور: رجال ونساء، فإن خوارزميات تعلم الآلة ستستمر في محاولة بناء مجموعة من النماذج (النمذجة) يستطيع من خلالها الحاسب أن يتنبأ أو يتوقع الجنس من الصورة المعطاة. وتكون مهمة وحدة التقييم اختيار أفضل نموذج من هذه النماذج المستخرجة.

١ , ١ نمذجة اللغة

لكن ما المقصود بنمذجة اللغة حاسوبيًا؟ يمكن للحاسب الآلي عبر خوارزميات الذكاء الاصطناعي بناء تمثيل معين للغة وذلك لاستخدامه في تطبيقات لاحقة. فمثلا، أحد التطبيقات المشهورة والمستخدمة بكثرة في الهواتف المتنقلة هي تطبيق لوحة المفاتيح الذكية؛ والتي تتيح للمستخدم عند كتابة كلمة، اختيار كلمة تالية لها. فمثلا، عند كتابة كلمة «السلام» يتيح الحاسب عدة اختيارات مثل «عليكم» أو «عليك». لكن كيف يمكن للحاسب «توقُّع» الكلمة التالية؟ لقد بنت خوارزمية الذكاء الاصطناعي(مثلا خوارزمية Skip-gram والتي تتنبأ بالسياق من الكلمة المعطاة) تمثيلاً لكل كلمة في اللغة يحدد موقعها من اللغة ككل، لذا فهو يستطيع أن يتنبأ بأقرب الكلمات اللاحقة لها.

وحينما نريد نمذجة لغة ما كاللغة العربية، فإن المهمة تكون عادة أصعب وأعقد؛ ذلك أن تحليل اللغة عادة ما يصحبها غموض في مستويات لغوية عدة كالصوت والصرف والنحو والمعجم. فمثلا الضمير في قولك: «قال زيد أنه مريض» غير معلوم؛ فقد يكون المريضُ زيدًا أو رجلا مقصودًا آخر. مثالٌ آخر فيما يروى عن الرسول –صلى الله عليه

وسلم- قوله: «نحن من ماء»، فتوهّم المخاطَب أنهم من ماء العراق. والغموض معلوم في النص العربي المكتوب، خصوصًا عند غياب التشكيل أو الترقيم أو الهمزات. وقد عمدوا قديمًا إلى الإعجام (إضافة النقاط إلى الحروف) من أجل إزالة جزء من الغموض (ومثله التشكيل). إلى جانب الغموض، هناك سبب آخر لصعوبة نمذجة اللغة، ألا وهو أن المتحدثون قد لا يلتزمون بجمل صحيحة نحويًا ودلاليًا وإملائيًا. والأمثلة على ذلك كثيرة، مثل الأخطاء الإعرابية والإملائية وأحيانًا الدلالية أو المعجمية؛ فقد يستخدم مفردة ليريد به مفردةً أخرى.

٢, ١ نمذجة الكلمة العربية

يهدف هذا الفصل إلى بناء مقدمة لغير المتخصص في كيفية نمذجة جزء محدد من اللغة: «الكلمة العربية الفصيحة المكتوبة». وبذلك تخرج اللغات غير العربية، وكذلك العامية التي لا يوجد لها نظام كتابي معياري. كما تخرج المهام التي تعنى بالفقرات أو النص كاملا، كتلخيص النص، أو استخراج موضوع الفقرة. كما يخرج من ذلك أي دراسة للكلمة المنطوقة والصعوبات التي قد تواجه الحاسب مثلا في تمييز الكلمة المنطوقة. وعند إيراد لفظة «الكلمة» فإن المقصود هي الكلمة المكتوبة (مجموعة الحروف التي تحدها مسافتان) لا الكلمة النحوية (مثل الضمير المتصل).

تناقش الورقة أيضًا آلية وصعوبات خوارزميات الذكاء الاصطناعي التي تهدف للقيام بمهام كثيرة متعلقة بالكلمة العربية تحديدًا: مثل تعيين قسم الكلام للكلمة، تعيين نوع الكلمة إذا كانت علمًا، تعيين نوع الاسم من حيث الجمع أو التثنية أو الإفراد وغير ذلك.

ولتوضيح المقصود من المهام، يمكننا دراسة الحديث الشريف التالي:

«لا يؤمن أحدكم حتى يكون هواه تبعًا لما جئت به» (الأربعين النووية)

فقد نبني عدة نماذج تدرس الكلمة:

- تعيين قسم الكلمة: لا/ حرف_نفي يؤمن/ فعل .. إلخ.
- تقسيم أجزاء الكلمة: أحدكم/ أحد+كم هواه/ هوا+ه لما/ ل+ما .. إلخ.

- التعرف على أصل الكلمة: يؤمن/ آمن أحدكم/ أحد يكون/ كان هواه/ هوى .. إلخ.
- تشكيل الكلمة: يؤمن/ يُؤْمِنُ أحدكم/ أَحَدُكُمْ .. إلخ.
- التعرف على الكلمة معجميًا: يؤمن/ آمن_التصديق (وليس الانقياد والطاعة) هواه/ هوى_الميل (وليس العشق) .. إلخ.
- التعرف على عائد الضمير: أحدكم/ المخاطبين هواه/ أحدكم جئتُ/ المتكلم .. إلخ.
- التعرف على بعض الخصائص الصرفية (مثل جنس الاسم وعدده، نوع وإعراب الفعل): يؤمن/ مرفوع يكون/ منصوب.
- استخراج معاني ومرادفات وأضداد من القاموس الشبكي للكلمات Wordnet: يؤمن/ الإيمان،الدين-مرادف:يصدق،يتبع،ينقاد-ضد: يكذب، يجحد

هذه النماذج تمثل النمط اللغوي لدراسة وتحليل الكلمة؛ وهذه النماذج غالبًا تمثل مرآة للعلم اللغوي واللساني الذي تطور عبر السنين. في السنوات الأخيرة، ظهر وانتشر نمط آخر لنمذجة الكلمة بناء على نظرية التوزيع الدلالي semantic distribution، والتي تستند على نظرية فيرث(Firth, 1957) والتي يقول فيها إنه «يمكن التعرف على الكلمة من الكلمات المصاحبة لها في النص». وبناء على النظرية، أصبحت مهمة الحاسب التعرف على الكلمة (م) بناء على الكلمات المصاحبة لها (أو السياق) (خ). هذه المهمة مشابهة للسؤال التعليمي الذي يطرحه المعلم على متعلمي لغة ما ليكملوا الفراغ في جملة بكلمة مناسبة، وهم بذلك يقيسون مدى استيعابهم ليس فقط للكلمة وإنما للسياق كذلك، وللتناسب بين الكلمة والسياق. وكذا الحاسب، كلما كان أكثر دقة في اختيار الكلمة الأنسب، كان تمثيل الكلمة أو نمذجتها أكثر فائدة ونفعًا.

كلا النمطان (اللغوي والتوزيعي) يمكن استخدامهما لتمثيل الكلمة حاسوبيًا. والتمثيل الحاسوبي أساسي للقيام بعدد كبير من العمليات. إذ لا يمكننا إجراء العمليات الرياضية على الكلمة وهي في شكلها الخام، مثل قياس المسافة أو الفرق. فمثلا، لا يمكننا القول (بسهولة) إن كلمة «مسجد» هي مفرد كلمة «مساجد» ولا أن

كلمة «مسجد» هي أقرب لكلمة «صلاة» من كلمة «شمس» وذلك فقط من خلال معرفتنا بمجموعة الحروف أو الأصوات التي تمثل الكلمة. بعبارة أخرى، نريد أن نكون قادرين على استنباط معادلة تُخبرنا أن الفرق بين البعد بين «مسجد» و»صلاة» أكبر من البُعد بين «مسجد» و«شمس»:

مسجد – صلاة > مسجد – شمس

هذه العمليات الرياضية قد تكون ضرورية في البحث. فيمكننا مثلاً عند البحث عن الصلاة، إظهار نتائج مثل أقرب المساجد إلى الباحث. عبر تمثيل الكلمات الثلاث بالنمط اللغوي أو التوزيعي أو بالإثنين معًا، فنستطيع المقارنة بين الصفات المستخرجة لكلمات البحث.

وهذه النماذج بمجموعها (أو بشكلها الرياضي) تعتبر مدخلاً مهمًا لتطبيقات أكثر تعقيدًا مثل استخراج المعلومات أو الترجمة الآلية أو تحليل الخطاب أو تلخيص النص أو توقع الكلمة التالية في لوحة المفاتيح الذكية، كما أسلفنا من قبل.

٢- صعوبات نمذجة الكلمة العربية

تواجه خوارزميات الذكاء الاصطناعي المصممة لتحليل ونمذجة الكلمة العربية صعوبات عدة. فاللغة العربية تصنف ضمن أغنى اللغات صرفيًا، والنظام الصرفي فيها ليس خطيًا، كما أن مستويات الغموض فيها عالية بسبب النقص المعتاد في اتباع النظام الكتابي (إهمال الهمزات والتشكيل مثلا). هذه العناصر الثلاثة تشكل أهم المصادر لصعوبة تحليل الكلمة العربية.

فاللغة العربية هي لغة غنية صرفيًا (Morphologically Rich Language). وهذا الغنى جعل التفاعل بين الصرف (دراسة بنية الكلمات) والنحو (دراسة علاقات الكلمات في الجملة) أكثر تعقيدًا. وكلما كانت اللغة أغنى صرفيًا، كانت الجملة تمتاز بمرونة أعلى في صفاتها، كالمرونة في ترتيب الكلمات (Tsarfaty et al., 2010). فمثلا، يمكننا قول: ضرب محمدٌ خالدًا، وضرب خالدًا محمدٌ. فوجود علامة الإعراب (الخصيصة الصرفية) سمحت بتقديم المفعول على الفاعل. وهذا قد يفسر لنا أهمية علامة الإعراب في النظرية النحوية التقليدية. هذه المرونة تجعل من خوارزميات الذكاء

الاصطناعي أقل قدرة على نمذجة اللغة (بالمقارنة مع لغة أقل مرونة كاللغة الإنجليزية) مع افتراض أن الخوارزميتين أُعطيتا نفس القدر من الأمثلة للتمرين (Heintz, 2014).

وهذا الغنى الصرفي يجعل الكلمة العربية نفسها كثيرة الاشتقاقات والالتصاقات، وذلك لترميز هذه الخصائص الصرفية. فعادة ما تحتوي الكلمة العربية الواحدة على مجموعة من الالتصاقات التي تبين خصائصها الصرفية، مثل «ـون» لبيان الجمع والمذكر، و«نـ» في الفعل المضارع لبيان المتكلم والجمع، و«تـ» كذلك لبيان إما المؤنث أو المخاطب كما في «أنت تصوم كثيرًا» و«فاطمة تصوم». وهذه الالتصاقات ليس مميزة بعلامة معينة مما يصعب تمييز اللواصق ويجعل الكلمة غامضة ومبهمة. فليس من السهولة بمكان تمييز الفاء الملتصقة في بداية الكلمة (حرف الاستئناف والعطف)، كما في الكلمة «فهم». فقد تكون الفاء ملتصقة فتكون مع الضمير المنفصل «فـ ـهُمْ» أو غير ملتصقة فيكون الفعل «فَهِم». إضافة لذلك، بعض العمليات الالتصاقية تغير في شكل الكلمة؛ أي أن الكلمة بعد الالتصاق تتغير طبقًا لقواعد صعبة النمذجة حاسوبيًا دون النظر في النظام الصرفي الغير خطي للغة العربية، مثل التصاق الكلمات معلولة الآخر بلاحقة: يَدْعُو – يَدْعُون، قالَ – قُلْت.

والخصائص الصرفية ليست دائمًا التصاقية، واللغة العربية –بالإضافة إلى العبرية– تتميز بكونها لغة سامية ذات نظام غير خطي وتبنى الكلمة فيها بناء على الوزن والجذر. وهذا جعل بعض الخصائص الصرفية صعبة التعلم من قبل الخوارزميات الحاسوبية. فمثلا، التعرف على الجمع في جموع التكسير ليس التصاقيا، وإنما اشتقاقيا مبني على وزن معين. وكذلك خوارزميات التعرف على جذر أو أصل الكلمة (Stemming) عادة ما تكون النتائج فيها ليست مثالية.

ونظرًا لأن الكلمة العربية كثيرة الاشتقاقات والالتصاقات، فإن عدد الأشكال المحتملة للكلمة الواحدة عالٍ جدًا، مما يؤدي بالضرورة إلى زيادة حجم وتباعد (sparseness) المفردات التي تخزن في النظام (ونقصد بها أشكال كل الكلمات العربية). وهذا التباعد يجعل من احتماليات توافق شكل كلمة معينة في النظام مع شكل كلمة أخرى محدودًا وقليلًا؛ مما يؤدي إلى تقليل كفاءة خوارزميات الذكاء الاصطناعي. ولتقليل آثار هذه المشكلة، يعمد كثير من الباحثين إلى تجزئة الكلمة آليًا إلى أجزائها

الرئيسية، فتصبح كلمة «ساعدوني» -مثلاً- مكونة من ثلاثة أو أربعة أجزاء حسب معايير التجزئة المستخدمة. وهذه التجزئة الآلية قد أثبتت فاعليتها مثلاً في تطبيقات الترجمة الآلية (Habash & Sadat, 2006) لكنها لا تخلو من أخطاء تؤثر في ما يلحق من خوارزميات.

ولأن الكلمة العربية غنية صرفيًا واشتقاقيًا، فإن عدد الخصائص المستخرجة للكلمة الواحدة أكبر من غيرها من اللغات مثل اللغة الإنجليزية. ففي حين يكفي لوصف كلمة إنجليزية أن ترمز برمز tag ضمن مجموعة رموز تتراوح بين ٣٠ إلى ٥٠ رمزًا، فإن مجموعة الرموز المحتملة لوصف الكلمة العربية تتجاوز ذلك بكثير وعادة ما تكون فوق المائة رمز (Habash, 2010). ففي حين يكفي أن نَصِفَ الاسم في اللغة الإنجليزية إما باسم_مفرد أو اسم_جمع، فإن الكلمة العربية «فرسٌ» قد تُرمَّز برمز أكثر تعقيدًا مثل اسم_مفرد_مؤنث_مرفوع_نكرة. والحاجة إلى أيٍّ من الخصائص يعتمد بالدرجة الأولى على الهدف المرجو من نمذجة الكلمة؛ فقد يحتاج إلى أغلب الخصائص في حالات الترجمة الآلية -مثلا-، بينما يكفي فقط إرجاع الكلمة إلى الأصل في تطبيقات البحث والتقصي.

بقي أن نشير إلى السبب الأخير في صعوبة نمذجة الكلمة العربية. وهو تسرب كثير من المحددات اللغوية عند كتابة النص، مثل غياب التشكيل أو الهمزات. هذا التسرب يزيد بشكل كبير من غموض الكلمة العربية. ففي دراسة لغموض الكلمة في كتاب رياض الصالحين (Alosaimy & Atwell, 2018)، وقياسًا على مجموعة من القواميس الحاسوبية، كان عدد الاحتمالات التحليلية للكلمة الواحدة (التحليل يشمل أصل الكلمة وقسمها وثمان خصائص صرفية) يقفز من معدل ٤.٨٣ احتمال إلى ما يقرب من ١٧.٤٢ احتمال عند غياب التشكيل.

٣-خوارزميات الذكاء الاصطناعي في نمذجة الكلمة لغويًا

كما أسلفنا فإن الطرق التحليلية اللغوية تشمل مجموعة الخوارزميات التي تنحى إلى استخراج الخصائص اللغوية للكلمة، وبعكس الطرق التحليلية التوزيعية التي تعنى باستخراج موقع الكلمة والعلاقات بينها وبين الكلمات الأخرى. أحد أهم الفروقات بين هذين الصنفين هو أن النمط اللغوي عادة ما يكون موجهًا تصنيفيًا بعكس النمط

التوزيعي الذي يكون غير موجه ويهدف إلى بناء مضامين الكلمة word embeddings (كما سنرى في الفصل القادم).

نقصد بتعلم الآلة الموجَّه العملية التي تحتوي على مجموعة من البيانات للتدريب S، وهذه البيانات هي عبارة عن أزواج من المدخلات والمخرجات المطلوبة (x,z). أي أنه لابد من أمثلة معطاة يتم التدريب عليها. ويكون المدخل x عنصرًا في المجموعة X والمخرج z عنصرًا في المجموعة Z. كما يضاف إلى مجموعة البيانات للتدريب، مجموعة بيانات أخرى منفصلة للتقييم S' (وأحيانًا مجموعة ثالثة للتحقق validation set) وكلا المجموعتين S و S' يمكن اعتبارهما مستخرجتين بشكل مستقل من نفس المصدر $D_{x\times z}$ (distribution). والهدف الأساسي من الخوارزمية هي استخدام مجموعة التدريب لبناء نموذج، مع الحرص على تقليل الخطأ الناتج E في مجموعتي التدريب والتقييم بقدر الإمكان. وعادة في الخوارزميات ذات العوامل Parametric algorithms (مثل الشبكات العصبية) يكون تطوير الخوارزمية عبر التعديل في العوامل التي تؤثر في دالة حساب الفقدان (Loss function).

فمثلا، لنأخذ عملية تحديد قسم الكلمة POS tagging (ليكون إما اسمًا أو فعلاً أو حرفًا). يمكننا أن نبني خوارزمية موجّهةً عبر إعطائها مجموعة تدريبية (مثلا: نام/ فعل صالح/ اسم على/ حرف السرير/ اسم...) ومجموعة أخرى لتقييم الخوارزمية (مثلا: أتى/ فعل خالد/ اسم من/ حرف السوق/ اسم..). لاحظ أننا نفترض أن كلا المجموعتين أتيا من نفس المصدر ولهما إذن نفس الخصائص التوزيعية (عدد متقارب لعدد الأسماء، نفس اللغة، ... إلخ). فلو افترضنا أن الخوارزمية بنت النموذج التالي (بعد النظر إلى مجموعة التدريب): إذا كانت الكلمة تبدأ بحرف النون فإنها فعلٌ، وإذا كانت تبدأ بحرف العين فإنها حرف، وما عداهما فهو اسم. سيصبح ناتج الخوارزمية ١٠٠٪ عند قياسها على المجموعة التدريبية، ولكن ٥٠٪ في المجموعة التقييمية. ولذا ستحاول الخوارزمية تطوير عملها، ربما عبر الذهاب إلى الحرف التالي، أو بأخذ أول حرفين بالاعتبار أو غير ذلك من الطرق.

يعيب تعلم الآلة الموجه حاجته إلى أمثلة للتدريب أي إلى «تعليم» أو «توسيم» البيانات annotation التي عادة ما تكون مجهدة ومكلفة ماديًا. ولكن بالمقابل، فإن

الخوارزميات الموجهة عادة ما تكون أفضل أداءً من نظيرتها الغير موجهة والتي لا تتطلب تحديد البيانات المطلوبة(Albared, Omar, & Ab Aziz, 2009).

عادة ما يكون تحليل الكلمة عبر بناء نموذج موجّه تصنيفي Classification Problem؛ أي أن الهدف الأساسي للنموذج هو القدرة على تحديد صنف معين بين عدة أصناف معروفة ومحددة سلفًا. وأحد أشهر النماذج وأكثرها نفعًا هو تصنيف قسم الكلمة POS tag، كالمثال السابق. إلا أن التصنيف –حتى وإن كان معياريًا– فإنه من صنع الإنسان وبناء على خبرته اللغوية ولذا فهو أحيانًا يفشل في الحالات الحدية borderline cases. ولذا فإن الباحثين يختلفون اختلافًا كبيرًا في تحديد الأصناف في أقسام الكلمة. فعند سيبويه، أن الكلمة اسم كفرس وحائط، أو فعل يدل على الحدث أو ما عدا ذلك وسماه الحرف. ولكن كثيراً من الباحثين اللغويين لم يرتأِ هذا التقسيم (كتمام حسان –رحمه الله– وتلميذه) وجعلها سبعة أقسام بناء على المبنى والمعنى: الاسم، والفعل، والضمير، والأداة، والصفة، والخالفة، والظرف. وقد أورد بعض الحالات الحدية، مثل اسم الفاعل (الذي يعمل عمل الفعل functional morphology، وله مبنى الاسم وصفاته form morphology) (الساقي، ١٩٧٥).

وعند بناء النموذج الموجه التصنيفي للغة ما، فإن النموذج عادة ما يأخذ التسلسل بين المدخلات في الاعتبار. ففي المثال السابق، يمكننا اعتبار التسلسل (مثال: «كل ما يلي الحرف فهو اسم» أو «الفعل يكون في بداية الجملة») ليزيد من دقة النموذج المستخرج. هذه الخاصية موجودة عادة في الخوارزميات التي تأخذ الوقت بالاعتبار، مثل تحويل الصوت إلى كلام، فليس من المنطقي اعتبار كل ثانية من الصوت جزءا مستقلا دون الأخذ بالاعتبار ما سبق من الثواني. وبهذا أصبحت المشكلة محددة أكثر ويمكننا تسميتها: الخوارزميات الموجهة التصنيفية لسلاسل البيانات.

هذه الخوارزميات يمكن تطبيقها على كثير من المهام اللغوية التي تتدرج من المستوى الصوتي phonology (أو الكتابي orthography)، الصرفي morphology، النحويsyntax، وحتى الدلاليsemantic. ففي المستوى الكتابي، يمكننا بناء نموذج موجه تصنيفي لتشكيل الكلمة. وتكون سلاسل البيانات فيه هي الحروف (وأماكن المسافات)، والأصناف هي علامات التشكيل. وتكون مجموع البيانات التدريبية

والتقييمية مأخوذة من ذخيرة لغوية مشكلة بالكامل. وكذلك في المستوى الكتابي-الصرفي، كالتعرف على أجزاء الكلمة من سوابق ولواحق. يمكننا كذلك تصنيف كل حرف في الكلمة إلى سابق أو أصل أو لاحق. وفي المستوى الصرفي يمكننا كما أسلفنا استخراج قسم الكلمة أو الخصائص الصرفية كالعدد والجنس والإسناد (للمخاطب أو للمتكلم أو للغائب) وغيرها من الخصائص. وهنا قد تكون سلاسل البيانات هي الكلمات نفسها أو أجزاء الكلمة المستخرجة سابقًا. وكذلك في المستوى الإعرابي في بناء شجرة الإعراب الإسنادية Transition-based Dependency Parsing Tree، يمكننا بناء الشجرة عبر بناء نموذج يأخذ كلمات الجملة كلمة كلمة (أو جزءًا جزءًا) من اليمين إلى اليسار، ثم يقرر في كل عملية إما أن يسند الكلمة إلى اليسار reduce right، أو إلى اليمين reduce left، أو أن يأخذ الكلمة التالية shift.

كما يقول المثل الإنجليزي: «العفريت في التفاصيل». على الرغم من أن بناء نموذج تصنيفي موجه (مع افتراض وجود بيانات موسومة للتدريب) سهل نسبيًا، إلا أن الحصول على دقة عالية تتطلب الدخول بشكل أكبر في التفاصيل. فمثلا، تختلف الخوارزميات المستخدمة في التصنيف اختلافًا كبيرًا، ولابد من معرفة الخوارزمية الأنسب بناء على الاحتياج. وكذلك، لابد من دراسة الخصائص المستخدمة في التصنيف بحرص، فالمثال السابق في تصنيف أقسام الكلمة أخذ الحرف الأول كخاصية مناسبة للتصنيف، ولكنها كما نعرف ليست الخيار الأفضل، وذلك لأنها محدودة النظر على مجموعة التدريب وضعيفة في تعميم مجموعة التدريب للأمثلة الحقيقية الواقعية.

فمثلا، هل المراد من الخوارزمية التصنيفية أن تكون تمييزية (discriminative model) أم توليدية (generative model)؟ النماذج التمييزية (مثل خوارزمية SVM والشبكات العصبية) تستخرج الاحتمال الشرطي للصنف بناء على البيانات المعطاة $p(c|d)$، ولكن النماذج التوليدية (مثل نموذج ماركوف Hidden Markov Model) تستخرج احتمالات الصنف والبيانات $p(c,d)$؛ أي تبني نموذجًا لكيفية توليد البيانات والأصناف ثم تصنف البيانات الجديدة لاحقًا بناء على نمذجتها للبيانات والأصناف. تتميز النماذج التمييزية بأنها أعلى دقة وأسرع تدريبًا وأسهل في دمج خصائص مختلفة، ولكنها في المقابل تحتاج إلى بيانات أكثر ولا تستطيع توليد بيانات شبيهة (أي لا يمكنها توليد كلمة شبيهة بناء على مجموعة من الخصائص)(Ng & Jordan, 2002).

كما أن الخوارزميات نفسها تختلف في أدائها. فعند تصنيف سلاسل البيانات، لابد من اختيار خوارزمية تأخذ بالاعتبار التسلسل مثل الشبكات العصبية التكرارية (مقابل الشبكات العصبية الأصلية). كما أن طريقة الربط بين النماذج المختلفة للكلمة (مثل التجزئة وتوسيم قسم الكلام) مؤثر على النتيجة. فمن البديهي، أن كل مهمة لاحقة تعتمد على نتائج المهمة السابقة، مما يؤدي إلى أن الأخطاء المولدة في المهام السابقة ستؤثر سلبًا على المهام اللاحقة. ولهذا السبب عمد بعض الباحثين إلى بناء نماذج تصنف عمليتين في الآن ذاته (Kudo & Matsumoto, 2001) الذي أثبت نجاعته في نمذجة الكلمة العربية تحديدًا (Algahtani & McNaught, 2015) . وفي السنوات الأخيرة، انتشر مفهوم النموذج المتكامل End–to–end model بفضل التقدم في مجال التعلم العميق في الشبكات العصبية، والذي يتيح تدريب مجموعة نماذج مختلفة للكلمة (مع إمكانية قياس دقة كل نموذج بشكل مستقل). هذا المفهوم أثبت مثلا أن تعلم الآلة للشجرة الإعرابية مفيد في زيادة دقة تصنيف قسم الكلام (Zhang, Li, Barzilay, & Darwish, 2015).

بالإضافة إلى اختيار الخوارزمية الأنسب، فإن البيانات نفسها واستخراج الخصائص منها تلعب دورًا كبيرًا في دقة النمذجة. ومن أشهر مصادر البيانات لتعلم الآلة النمط اللغوي: البنوك الإعرابية الشجرية Treebanks؛ وهي تحتوي على كمٍّ كبير من التوسيمات على مستوى الكلمة: معجميًا (أصل الكلمة)، وصرفيًا (قسم الكلمة والخصائص الصرفية بالإضافة إلى بيان اللواصق وأجزاء الكلمة)، وإعرابيًا (العلاقات بين الكلمات أو تركيبة الجملة).

أشهر البنوك الإعرابية البنك الشجري العربي من بنسلفينيا (PATB) (Maamouri & Bies, 2004)، الذي يحتوي على نصوص إخبارية (ما يقارب ٧٥٠ ألف كلمة) باللغة العربية المعاصرة مجزأة ومسومة بتشكيل الكلمة وأصل الكلمة (تحديدًا المدخل المعجمي Lemma) وقسم الكلمة والخصائص الصرفية طبقًا لمجموعة أصناف تيم بكولتر بالإضافة إلى شجرة الإعراب لكل جملة (contingency treebank). كما أن هناك مصادر أخرى مثل تلك المتاحات ضمن موقع شجرات الإعراب العالمي[1] وجدول ١ يسرد أشهر البيانات المتاحة الموسومة.

1- http://universaldepdencies.com

الاسم	متاح	الأصناف	عدد الكلمات	نوع النص	مرجع
جامعة الإمام	لا	تجزيء النص دون توسيمه	٧٧ ألف	قرآني	(Elhadj, Al-Sughaiyer, Khorsi, & Alansari, 2010)
سلمى	نعم	أصنافه الخاصة	ألف	قرآني	(Sawalha, 2011)
الذخيرة الدينية	عند الطلب	بنسلفينيا	٢٧ ألف	قرآن، سنة، والفلسفة	(Mohamed, 2012)
الذخيرة القرآنية	نعم	أصنافه الخاصة	٧٧ ألف	قرآني	(Dukes, Atwell, & Habash, 2013)
المصحف	نعم	أصنافه الخاصة (محلل الخليل)	٧٧ ألف	قرآني	(Zeroual & Lakhouaja, 2016)
ذخيرة خوجة	عند الطلب	أصنافه الخاصة	٥٠ ألف	أخبار	(Khoja, 2001)
بنسلفينيا	باشتراك مدفوع	البنك الشجري من بنسلفينيا (محلل بكولتر)	٧٥٠ ألف	أخبار	(Maamouri, Bies, Buckwalter, Jin, & Mekki, 2005)
براغ	متاح	أصنافه الخاصة(محلل إكسير)	١١٣ ألف	أخبار	(Hajic, Smrz, Zemánek, Šnaidauf, & others, 2004)
كاتب (كولومبيا)	باشتراك مدفوع	أصنافه الخاصة قليلة	مليون*	أخبار	(Habash & Roth, 2009)
نيملار	مدفوع	أصنافه الخاصة	٥٠٠ ألف	١٣ مجالا	(Yaseen et al., 2006)
أقمار	نعم	كاتب	٣٦ ألف	ويكيبيديا	(Schneider, Mohit, Oflazer, & Smith, 2012)

جدول ١ : قائمة بالبيانات العربية الموسومة بالتحليل اللغوي.

كل هذه البيانات وغيرها (خصوصًا البنك PATB) مستخدمة بكثرة في تدريب خوارزميات متنوعة لنمذجة اللغة ومن أشهرها المحللات الصرفية. ولأن تركيز هذه الورقة على الخوارزميات لا البيانات (التي هي خارج إطار هذا البحث)، فإنه يكفينا أن نبين اختلافات ملحوظة يجب الانتباه لها عند تطوير أو أخذ خوارزمية بعين الاعتبار:

- البيانات ليست كلها متاحة للتحميل، فبعضها مجاني ومفتوح، وبعضها يمكن الحصول عليه مباشرة من الباحث، وبعضها لابد من اشتراك مدفوع في المؤسسة المانحة (مثل شجرة PATB).
- تختلف طريقة توسيم البيانات بشكل كبير جدًا. وكل اختلاف في مستوى لغوي أبسط يؤثر في المستويات اللاحقة. فمثلا، الاختلاف في طريقة تجزئة النص يؤدي إلى الاختلاف في أصناف قسم الكلمة مما يؤدي إلى الاختلاف في الشجرة الإعرابية.
- سلاسل البيانات عادة ما تكون موسومة برموز تصف هذه السلاسل، لكن تختلف البيانات في أين يكون توسيم «جزيئات» النصوص token، فقد تكون على مستوى الكلماتٍ أو أجزاء منها أو حتى الحروف المكونة لها.
- عند اعتبار الكلمات أساسًا للتوسيم، فإن تحديد الجزء الأصيل من الكلمة من غيره من اللواصق قد يكون مشكلا. مثل تحديد الجزء الأصيل في الكلمة: «معهم»، هل هو حرف الجر أو الضمير.
- عادة ما يبنى التوسيم الصرفي في البنوك الشجرية طبقًا لمحلل صرفي يسهل عملية التوسيم. فبدلا من أن تكون يدوية بالكامل، فإن الواسم يختار أحد التحليلات الصرفية المقترحة من المحلل. كما بإمكانه إضافة تحليل جديد إذا لم يجد مبتغاه. معرفة المحلل الصرفي وخصائصه مهمة، حيث إن تأثيره على البنك الشجري بالغ.
- لا ينبغي الاعتماد على الأصناف الموجودة في البنك الشجري عند بناء محلل ما. ولكن يجب قَصْر الأصناف على المطلوبة فقط لتحقيق الهدف النهائي.

١ , ٣ المحللات الصرفية

في هذا القسم، تناقش الورقة طرقًا مستخدمة في نمذجة اللغة (وتحديدًا في المحللات الصرفية) ونحلل أربعة من أشهر المحللات اللغوية العربية، وهي تحديدًا: مداميرا، أميرة، ستانفورد، وفراسة.

المحللات الصرفية نوعان: ترميزي (حيث يقرر المحلل التحليل الصرفي الأنسب للكلمة، انظر أشهرها في جدول ٣) ومعجمي (حيث يسرد المحلل التحليلات الممكنة للكلمة دون الاختيار أو التفضيل انظر أشهرها في جدول ٢). المحللات الصرفية الأربعة هي من الصنف الأول، وتجتمع كلها في أن التدريب قد تم على البنك الشجري من بنسلفينيا وعلى تصغير عدد الأصناف المستخدمة في البنك. ولكن كل محلل يختلف في تصميمه وطريقة نمذجة الكلمة والخصائص التي يقدمها ويتنبأ بها، انظر جدول ٣.

الخاصية	Buckwalter بكولتر	AlKhalil الخليل	Elixir إكسير	AraComLex أراكوملكس
صنف الكلمة	نعم	نعم	نعم	نعم
نوع الفعل	نعم	نعم	نعم	نعم
الإسناد	نعم	نعم	نعم	–
الجنس	نعم	نعم	نعم	نعم
العدد	نعم	نعم	نعم	نعم
التعدية	–	نعم	–	نعم
البناء للمجهول	نعم	نعم	نعم	نعم
التعريف	نعم	نعم	نعم	–
إعراب الفعل	نعم	نعم	نعم	–
إعراب الاسم	نعم	نعم	نعم	–
الوزن	–	نعم	نعم	–
الجذر	–	نعم	نعم	نعم
الجذع	نعم	نعم	نعم	–
المدخل المعجمي	نعم	نعم	–	–

الخاصية	**Buckwalter** بكولتر	**AlKhalil** الخليل	**Elixir** إكسير	**AraComLex** أراكوملكس
التشكيل	نعم	نعم	نعم	–
مصطلح إنجليزي	نعم	–	نعم	–
التجزي	نعم	نعم	نعم	نعم
نوع التوسيم	مستوى الكلمة	مستوى الكلمة	مستوى الجزء	مستوى الكلمة
مرجع	Buckwalter,) (2002	(Boudchiche, Mazroui, Bebah, Lakhouaja, & Boudlal, 2016)	(Smrz, 2007)	(Attia, 2006)

جدول ٢: خصائص أشهر المحللات الصرفية العربية المعجمية (التي لا تزيل الغموض).

يتميز محلل مداميرا في أن تحليله الصرفي يعتمد على محلل صرفي معجمي مضمَّن في داخله (نسخة مطورة من المحلل المشهور بكولتر). ففي بداية عمله، يقوم مداميرا بتوقع نتائج التصنيفات لأقسام الكلمة ومجموعة من الخصائص الصرفية، ثم يقوم بعدها بمقارنة النتائج مع نتائج المحلل المضمَّن واختيار الأنسب. كما يتميز بأن العملية التصنيفية تتم على مستوى الكلمة لا على مستوى الجزء، فلا يوجد تجزيء للكلمة قبل التحليل، ولكن نتائج التحليل تعيّن اللواصق للكلمة (تفترض أن للكلمة أربع سوابق، ولاحقة واحدة بحد أقصى). اعتماد مداميرا على المحلل المضمَّن مكنه من رفع مستوى النتائج، لكن هذا محدود في حال تعرّف المحلل المضمن على الكلمة. كما أن اعتماده على التحليل على مستوى الكلمة قلل من تأثير الأخطاء المتولدة عند التجزئة على مرحلة التوسيم الصرفي.

محلل أميرا اختار طريقة التجزئة والتوسيم في آن واحد وذلك على عبر تجزئة الكلمة إلى حروف ثم تعيين مكان ووسم الحرف ثم تجميع الكلمة بناء على مكان أحرفها. فمثلا الكلمة فَهُمْ ستصبح فَـ/ عطف_سابق١ ـهُـ/ ضمير_أساس ـمْ/ ضمير_أساس. أما محللا ستانفورد وفراسة فإنهما اعتمدا النظام الخطي حيث التجزئة تسبق التوسيم مع تميّز فراسة باعتماده على كثير من المعاجم والفهارس لتيسير عملية التحليل.

الاسم	مداميرا	ستانفورد	أميرة	فراسة
صنف الكلمة	نعم	نعم	نعم	نعم
نوع الفعل	نعم	نعم إلا للمجهول	نعم	-
المخاطب والمتكلم والغائب	نعم	-	نعم	-
الجنس	نعم	-	نعم	نعم للأسماء
العدد	نعم	مفرد وجمع فقط	نعم	نعم للأسماء
المعلوم والمجهول	نعم	نعم	نعم	-
التعريف	نعم	-	-	-
إعراب الفعل	نعم	-	-	-
إعراب الاسم	نعم	-	-	-

جدول ٣: خصائص أشهر المحللات الصرفية المرمزة

٤- نمذجة الكلمة توزيعيًا

نمذجة الكلمة لغويا كما في الفصول الماضية، أنتجت لنا قيما لغوية معينة لكل خاصية من الخصائص اللغوية للكلمة. فمثلا، أصبحت كلمة «ضارب» في الجملة: «كان ضارب الناس أجنبيا» ممثلة بالشكل التالي:

<الوزن=فاعل، قسم_الكلمة=اسم، التشكيل=ضارِب، الحالة=مرفوع، العدد=مفرد، وغيرها>

هذه النمذجة مبنية على تصنيف الإنسان والذي تطور عبر دراسة اللغة عبر العصور. في هذا القسم سندرس طريقة مختلفة لنمذجة الكلمة، تتجاوز التمثيل اللغوي، إلى استخراج تمثيل مبني على نصوص اللغة.

يعيب التمثيل اللغوي للكلمة أنه يعتبر تمثيلا تصنيفيا لا تمثيلا رقميا. فمثلا، لا يمكننا اعتبار المرفوع رقما كالواحد والمنصوب كالاثنين والمجرور كالثلاثة، وذلك لأن ذلك يقتضي أن هناك ترتيبًا معينًا بين القيم الثلاث. والتصنيف لا يقتضي أي ترتيب بين الأصناف (رغم أن الأصناف عادة ليست بنفس التباعد). هذا الأمر ينطبق على مفردات اللغة نفسها. فلا يمكننا تحويل كلمات اللغة إلى أرقام مباشرة ولحل هذا الأمر،

طور الباحثون طرقاً عديدة لتمثيل الكلمة كمتجه رقمي numerical vector. ما يميز المتجه الرقمي (كالأرقام) عن الصنف، أنه يتيح لنا القيام بعدد من العمليات الرياضية كحساب المسافة بين متجهين.

١ , ٤ التمثيل الكلاسيكي للكلمة

الخيار الأول هو تمثيل كل كلمة بمتجه طويل (طوله عدد مفردات اللغة) ذو قيم كلها أصفار ما عدا عمود واحد. يمكننا تسمية هذا المتجه بالمتجه ذو الرقم الواحد one–hot encoding. ولنضرب مثالاً على هذا المتجه: لو افترضنا أن عدد كلمات اللغة ثلاث كلمات: «فرس، حائط، حصان» فإن الكلمة الأولى تمثل بالمتجه <٠ , ٠ , ١> والثاني <١ , ٠ , ٠> وهكذا.

يعيب هذا الخيار أنه يفترض مسافة واحدة بين أي كلمتين في اللغة. ولكن الحقيقة هي أن الفرس أقرب إلى الحصان منه إلى الحائط. بإمكاننا للحصول على مثل هذا التمثيل، الاستناد إلى نظريات «دلالات التوزيع» distributional semantics. هذه النظرية تستند إلى القول بأن العناصر اللغوية ذات التوزيع نفسه لها المعنى نفسه.

أحد الأمثلة على التمثيل التوزيعي، التمثيل الكلاسيكي المبني على توارد الكلمات Co–occurrence. فإذا كان الفرس والحصان يشتركان في الكلمات المتواردة معهما، فإن لهما المعنى نفسه. فمثلا، قد نجد أن كلاهما يردان مع الكلمات: حذوة، سرج، إلخ فنستدل أن لهما معنى متقاربًا.

وفي التمثيل التوزيعي الكلاسيكي، يتم بناء مصفوفة مربعة طولها وعرضها عدد مفردات اللغة. وفي كل خانة يتم تعداد عدد المرات التي وردت فيها كلمة ما في سياق كلمة أخرى. وبذلك أصبح تمثيل كل كلمة هو الصف المستخرج من هذه المصفوفة. الجدول أدناه مثال على المصفوفة للمثال السابق.

	فرس	حائط	حصان
فرس	٠	١	٢
حائط	١	٠	٠
حصان	٢	٠	٠

جدول ٤ : مثال على مصفوفة توارد الكلمات. الخانة الرقمية في الزاوية العليا اليسرى تعني أن كلمة حصان وردت مرتان في سياق كلمة فرس. التمثيل المتجهي لكلمة فرس هو <٢ , ١ , ٠>.

مما يعيب مصفوفة توارد الكلمات هو حجمها الكبير، خصوصًا في اللغات الغنية صرفيًا، وفقرها من ناحية تمثيل الكلمة لتباعد المفردات sparseness، فكل أشكال الكلمات تدخل في عدد مفردات اللغة والتوارد بين هذه الأشكال أكثر الأحيان معدوم. وهذا يجعلها مكلفة وغير عملية عند تمرين كثير من الخوارزميات.

٢ , ٤ مضامين الكلمة

مؤخرًا، ظهر تمثيل حديث للكلمة سُمِّيَ بمضامين الكلمة word embedding يتميز بكونه عملي وسريع مع الحفاظ على تمثيل دلالي جيد للكلمة، وأشهر أداة تنتج هذه المضامين هي word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). وهذا التمثيل يعتمد على خوارزمية الذكاء الاصطناعي: الشبكات العصبية.

الناتج النهائي من خوارزميات مضامين الكلمة هو تمثيل كل كلمة من كلمات اللغة بمتجه رقمي طوله محدد (عادة ما يكون ٣٠٠ عنصر رقمي) بحيث تكون الكلمات ذات المعاني المتقاربة ذات ضرب نقطي للمتجهين يقترب من الواحد (أي أن الزاوية بين المتجهين تساوي أو تقترب من الصفر).

هذا الطريقة في تسهيل التعامل مع الكلمات عبر تحويلها إلى متجهات رقمية يمكن استخدامها كذلك في الأصناف اللغوية (أو أية قيم تصنيفية categorical data (وفي هذه الحالة تسمى الطريقة بناء المتجه الكثيف). فالمشهور هو استخدام المتجه ذو الرقم الواحد one-hot encoding لهذه الأنواع من البيانات. وهذا التمثيل لا يصلح خصوصًا عند وجود عدد كبير من الأصناف. فمثلا، نشر تشن ومانتق بحثاً بأن استخدام المتجه الكثيف (أو المضامين) سبب سرعة أكبر في إعراب الكلمات syntactic

parsing. كما نشر أحد الباحثين استخداما لمضامين قسم الكلمة (بدلا من قسم الكلمة نفسه) من أجل نمذجة الجمل والفقرات(Yu, 2016).

٣,٤ إنشاء مضامين الكلمة

الخوارزمية word2vec تتكون من عدة مراحل. في المرحلة الأولى، تجرد الخوارزمية كل المفردات الواردة في الذخيرة اللغوية. ثم تستخرج سياقات كل كلمة من نافذة محددة (مثلا الكلمتين المجاورتين للكلمة) فتبني منها جدولا يبين توارد الكلمات، كالجدول أدناه.

الكلمة المعنية	كلمة السياق	التوارد
حائط	البراق	١
حائط	يرتبط	١
...		
الإسراء	بقصة	١
الإسراء	والمعراج	١

جدول ٥: استخراج سياقات في العبارة: «بالنسبة للمسلمين، يرتبط حائط البراق بقصة الإسراء والمعراج» مع اعتبار السياق الكلمتين المجاورتين للكلمة.

ثم تقوم الخوارزمية ببناء مصفوفتين: مصفوفة مضامين الكلمة المعنية (عادة الكلمة الوسطى)، ومصفوفة مضامين كلمات السياق. كلا المصفوفتين لهما طولٌ بعدد مفردات اللغة، وعرض بحسب طول المتجه النهائي المطلوب (عادة ٣٠٠). كلا المصفوفتين ينشآن بأرقام عشوائية ابتدائية.

يتم التدريب عن طريق أخذ الكلمات كلمة كلمة. ولكل كلمة ينشأ جدول السياقات مع إضافة سياقات خاطئة سلبية negative sampling ذات توارد صفري.

هناك نمطان للخوارزمية: النمط الأول رزمة الكلمات continous bag of words (CBOW) وبها يُتنبأ بالكلمة المعنية من سياقها، والنمط الثاني skipgram يحاول أن يتنبأ بالسياق من كلمة معينة.

لو أخذنا النمط الثاني في الاعتبار، فإن الخوارزمية ستقوم بعد ذلك بضرب نقطي بين مضمون الكلمة المعنية مع مضامين كلمات السياق (أي أنها تحسب مقدار الزاوية بين هذه الكلمات). ومن ثم يتم حساب الخطأ في الناتج من الضرب (يجب أن يساوي واحدًا للسياقات الإيجابية وصفرًا للسياقات السلبية). بعد ذلك، ستقوم الخوارزمية بتعديل المضامين في كلا المصفوفتين لتقليل هذا الخطأ.

تستمر المضامين في التحسن بينما تعاد هذه العملية لكل كلمة في الذخيرة اللغوية، وأحيانا تعاد لعدة دورات على كامل الذخيرة epochs. الناتج النهائي من الخوارزمية هي مصفوفة الكلمة المعنية وفيها تمثيل لكل كلمة بشكل متجه رقمي.

٤ , ٤ تقييم مضامين الكلمة

بمجرد إنشاء هذه المتجهات يمكننا استخراج أقرب الكلمات لكلمة ما وفحص جودة التمثيل. وهناك طرق علمية لتقييم هذا التمثيل الرقمي للكلمة منها التناظر اللفظي، الاختيار الأمثل للكلمة ضمن سياق، والقاموس العكسي.

في التناظر اللفظي، يُعطى الحاسب كلمتين متناظرتين (دون تحديد سبب التناظر) ثم يُطلب منه الكلمة المناظرة لكلمة ثالثة. وكمثال، يعطى الكلمتان المتضادتان حار:بارد، ويطلب منه المناظر لكلمة أعلى (أسفل). هذه الطريقة تقيس جودة تمثيل العلاقات بين الكلمات ويتم تنفيذها عبر عمليات رياضية مثل الجمع والطرح.

لهذه الخوارزمية عدة عوامل تؤثر في عملية التدريب، منها: النمط المستخدم وطول نافذة الكلمات. النمط الأول أسرع تمرينًا وعادة ما يكون للذخيرة اللغوية الطويلة ويمثل الكلمات المتكررة بشكل أفضل. أما النمط الثاني فيتميز بأنه يمثل الكلمات النادرة بشكل أفضل. طول النافذة القصير عادة ما يقرّب بين الكلمات القابلة للتبديل بين بعضها البعض، وطول النافذة الطويل عادة ما يقرّب الكلمات المتعلقة ببعضها.

٥ , ٤ تطور مضامين الكلمة

كانت هذه الخوارزمية بمثابة الشرارة لكثيرٍ من الأبحاث التي تعالج جوانب القصور فيها وتطور نماذج أكثر دقة في تمثيل الكلمة.

أحد جوانب القصور في الخوارزمية أنها لا تأخذ الترتيب في كلمات السياق ولا

تأخذ تركيبة الكلمة بالاعتبار عند التدريب. وهاتان الخاصيتان تحديدًا مهمتان في اللغة العربية، حيث إن وزن الكلمة يلعب دورًا مهمًا في معنى الكلمة، كما أن بعض السوابق واللواحق مؤثرة في المعنى (مثل التاء المربوطة).

من العيوب المشتهرة عن الخوارزمية أنها لا تفرق بين المتشابهات اللفظية homographs. فمثلا المتجه الناتج لكلمة «عين» سيكون مختلا، وذلك لأن الكلمة لها عدة معانٍ تأتي في سياقات متباينة. ففي آلية عمل الخوارزمية، تقوم الخوارزمية بتصحيح معامل الخطأ كل مرة ترد فيها الكلمة بشيء مناسب للكلمات المصاحبة (في السياق). ولكن الكلمة لها سياقات متباينة مما يؤدي إلى إنتاج متجه رقمي وسطي (حسب عدد مرات تكرار كل معنى).

ومن العيوب أيضًا، محدودية الخوارزمية على الكلمات التي وردت في الذخيرة التي تم التدريب عليها. فهي ليست تعميميَّة بحيث يمكنها التنبؤ بالكلمات التي لم ترد سابقًا في الذخيرة.

من أجل ذلك قام باحثون بتطوير عدة نماذج معدلة على الصيغة الأساسية. من ذلك حزمة fastText التي طورها باحثون في شركة فيسبوك، والتي لا تكتفي بتمثيل الكلمات بشكلها النهائي ولكن تأخذ أبعاض الكلمة (subword) بعين الاعتبار. فتمثيل الكلمة الناتج عن هذه الحزمة هو مبني على كل أبعاضها الكتابية سواء الحروف أو الأبعاض الثنائية bigram أو الثلاثية أو حتى أكثر من ذلك. فمثلا، التمثيل الخاص بكلمة «خيل» سيتكون من مجموع تمثيل كل حرف من حروفها الثلاثة، بالإضافة إلى تمثيل بَعْضَيْها «خي» و «يل». هذه الحزمة أفضل من سابقتها كونها تأخذ الصرف في الاعتبار (واللغة العربية غنية صرفيًا)، ويمكنها التنبؤ بكلمات لم ترد من قبل (واللغة العربية كثيرة الاشتقاقات والالتصاقات).

لكن التطور في هذا المجال لم يتوقف. فقد تنبه الباحثون إلى أن الخوارزمية لا تأخذ السياق بالاعتبار. نعم، هي تولي الكلمات المصاحبة للكلمة المعنية اهتماما، لكن هذه الكلمات استبعدت من السياق الكامل للفقرة أو حتى المقالة. وعالجت هذا القصور خوارزميات تستخدم طبقات ذات خصائص تذكريّة لسلاسل البيانات وأشهرها طبقة الذاكرة قصيرة المدى المطولة LSTM. كما طورت أكثر عبر خاصية التركيز attention

والتي تعطي اهتمامًا أكبر للكلمات ذات التأثير الأكبر عند التنبؤ بالكلمة التالية. فمثلا في الجملة التالية: «ويحب الأطفال اللون الأخضر والأحمر و___»، فإن التركيز من أجل التنبؤ بالكلمة الناقصة سيكون منصبا أكثر على كلمتي «الأخضر» و«والأحمر» كونهما مؤثران على الاختيار. وذاع صيت النماذج المستخرجة بهذه الطرق (مثل نماذج BERT (Devlin, Chang, Lee, & Toutanova, 2018) وGPT-2 (Radford et al., 2019)) التي أظهرت جودتها في توليد فقرات كاملة بشكل آلي متوافقة بشكل كبير مع السياق لفقرة مكتوبة معطاة.

٤,٦ تطبيقات مضامين الكلمة

أخيرًا، هذا التمثيل الرقمي للكلمات مفيد في عدد كبير من التطبيقات اللاحقة (downstream tasks). فكما أسلفنا هو مستخدم الآن في كثير من لوحات المفاتيح الذكية للتنبؤ بالكلمة التالية. كما أنه مفيد حتى في تصنيف الخصائص اللغوية بالنمط الأول (مثل تصنيف قسم الكلام).

وبعد تطور خوارزميات استخراج مضامين الكلمة (حتى أنها سميت بنمذجة اللغة Language Modeling إذ إنها لم تعد مقتصرة على الكلمة فقط وإنما تعتبر سياق الجمل والفقرات)، أصبحت شبه أساسية وأولية لكثير من مهام مجال معالجة اللغات الطبيعية. فمثلا، في النموذج BERT، رفع استخدام مضامين الكلمة السياقية من جودة إحدى عشرة مهمة من مهام معالجة اللغات الطبيعية. وصار على المستخدم الذي يريد تحليل المشاعر في نص معين أن لا يبدأ من الصفر، وإنما يبدأ من نموذج لمضامين الكلمات (قد سبق تدريبه) ثم يبني نموذجه لتحليل المشاعر فوقه بكل سهولة.

٥- خاتمة

تطورت خوارزميات الذكاء الاصطناعي المستخدمة لتحليل ونمذجة الكلمة تطورا كبيرا في السنوات الأخيرة، والتي تعد خطوة أساسية مهمة لكثير من تطبيقات معالجة اللغة الطبيعية. فلا شك أن معرفة المعاني والمباني للكلمات (اللبنات الأساسية للكلام) ضروري من أجل فهم الكلام أو إنتاجه آليا. ولذا تعددت الطرق التي تفسر وتمثل الكلمات وتبني جسرا ليسهل فهمه من قبل الحاسب الآلي الذي لغته لا تتجاوز الأرقام.

صنفت هذه الورقة الأبحاث في هذا المجال إلى نمطين: النمط اللغوي المبني على الدراسات اللغوية للكلمة مثل التحليل الصرفي والنمط التوزيعي الذي يستنتج المعاني للكلمة بناء على سياقاتها وتوزيعها في النص.

شرحت هذه الورقة كثيرا من الخوارزميات المستخدمة في تحليل ونمذجة الكلمة في كلا النمطين، مع تبيان جوانب القصور عند بعضها وخصوصا عند تطبيقها لنمذجة الكلمة العربية. كما قارنت بين النمطين وبينت أنهما مكملان لبعضها، إذ يمكن الاستفادة من النمط التوزيعي كخطوة أولى لتحليل كثير من المهام في النمط اللغوي.

بقي أن نختم بأن جزءاً كبيراً من التطور في النمط التوزيعي منصبّ بشكل كبير على اللغة الإنجليزية، وكثير من النماذج المستخرجة هي لذات اللغة، والأبحاث في اللغة العربية متأخرة في تجربة مثل هذه الخوارزميات وقياس جودتها على اللغة العربية تحديدا، حيث إن للغة العربية مشاكلها الخاصة. فمثلا، تشكيل الكلمة هي ميزة شبه فريدة للغة العربية وعدم التشكيل ليس خطأ إملائيا، وإنما وجوده اختياري ومتفاوت وتقديري. ويعمد كثير من الباحثين للغة العربية إلى إزالته منعا للزيادة في تباعد الكلمات sparseness. ولكن هذه الإزالة تزيد من غموض الكلمة (والتي بسببها يعمد الكاتب كثيرا إلى كتابة التشكيل أو الشدة). وعليه فيجب البحث عن إمكانية بناء خوارزمية ذكية لبناء المضامين تستطيع تمييز الحركات وتستفيد منها دون أن تؤثر على تباعد الكلمات.

المراجع

الساقي، ف. (١٩٧٥). أقسام الكلام العربي من حيث الشكل والوظيفة. مكتبة الخانجي بالقاهرة.

Albared, M., Omar, N., & Ab Aziz, M. J. (2009). Arabic part of speech disambiguation: A survey. International Review on Computers and Software, 4(5), 517–532.

Algahtani, S., & McNaught, J. (2015). Joint Arabic Segmentation and Part-Of-Speech Tagging. In Second Workshop on Arabic Natural Language Processing (p. 108).

Alosaimy, A., & Atwell, E. (2018). Diacritization of a Highly Cited Text: A Classical Arabic Book as a Case. In 2nd IEEE International Workshop on Arabic and derived Script Analysis and Recognition (ASAR 2018). London, UK.

Attia, M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In The Challenge of Arabic for NLP/MT Conference. London: The British Computer Society.

Boudchiche, M., Mazroui, A., Bebah, M., Lakhouaja, A., & Boudlal, A. (2016). AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. Journal of King Saud University-Computer and Information Sciences.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.0(Mlm).

Dukes, K., Atwell, E., & Habash, N. (2013). Supervised collaboration for syntactic annotation of Quranic Arabic. Language Resources and Evaluation. https://doi.org/10.1007/s10579-011-9167-7

Elhadj, Y., Al-Sughaiyer, I. A., Khorsi, A., & Alansari, A. (2010). The Morphological Analysis of the Holy Qur'an: A Database of the Entire Quranic Text (Arabic). International Journal of Computer Science and Engineering in Arabic, 3(1).

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (Special Volume of the Philological Society), 1952–59, 1–32.

Habash, N. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Habash, N., & Roth, R. M. (2009). CATiB: the Columbia Arabic Treebank. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 221–224). Suntec, Singapore.

Habash, N., & Sadat, F. (2006). Arabic Preprocessing Schemes for Statistical Machine Translation. In Proceedings of the Human Language Technology Conference of the NAACL. New York City, US.

Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., & others. (2004). Prague Arabic dependency treebank: Development in data and tools. Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools, 110–117.

Heintz, I. (2014). Language Modeling. In I. Zitouni (Ed.), Natural Language Processing of Semitic Languages (pp. 161–196). Springer. https://doi.org/10.1007/978-3-642-45358-8

Khoja, S. (2001). APT: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at NAACL (pp. 20–25). Pittsburgh, PA, USA.

Kudo, T., & Matsumoto, Y. (2001). Chunking with support vector machines. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics. Pittsburgh, Pennsylvania. https://doi.org/10.3115/1073336.1073361

Maamouri, M., & Bies, A. (2004). Developing an Arabic treebank: methods, guidelines, procedures, and tools. Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages, 2–9.

Maamouri, M., Bies, A., Buckwalter, T., Jin, H., & Mekki, W. (2005). Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis) LDC2005T20.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems 26 (NIPS 2013) (pp. 1–9). Lake Tahoe, USA.

Mitchell, T. M. (1997). Does Machine Learning Really Work? AI Magazine, 18(3), 71–83.

Mohamed, E. (2012). Morphological Segmentation and Part of Speech Tagging for Religious Arabic. In Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4) (pp. 65–71). San Diego, USA.

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative Vs. Generative Classifiers: a Comparison of Logistic Regression and Naive Bayes. In Advances in neural information processing systems (pp. 841–848).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Sawalha, M. (2011). Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora, PhD thesis. University of Leeds.

Schneider, N., Mohit, B., Oflazer, K., & Smith, N. (2012). Coarse lexical semantic annotation with supersenses: an Arabic case study. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 253–258). Jeju Island, Korea.

Smrz, O. (2007). Functional Arabic Morphology. Formal System and Implementation, PhD thesis. The Prague Bulletin of Mathematical Linguistics. Charles University in Prague.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Versley, Y., Candito, M., … Tounsi, L. (2010). Statistical Parsing of Morphologically Rich Languages (SPMRL): What, How and Whither. In Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages (pp. 1–12). Los Angeles, CA, USA: Association for Computational Linguistics.

Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., … Ragheb, A. (2006). Building Annotated Written and Spoken Arabic LR’s in NEMLAR Project. In LREC: Proceedings of the International Conference on Language Resources and Evaluation (pp. 533–538). Genoa, Italy.

Yu, D. J. (2016). Part-Of-Speech Tag Embedding for Modeling Sentences and Documents, Master Thesis. University of California, Los Angeles.

Zeroual, I., & Lakhouaja, A. (2016). A New Quranic Corpus Rich in Morphosyntactical Information. International Journal of Speech Technology, 1–8.

Zhang, Y., Li, C., Barzilay, R., & Darwish, K. (2015). Randomized Greedy Inference for Joint Segmentation, POS Tagging and Dependency Parsing. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 42–52). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1005

الفصل الخامس

تقنيات الذكاء الاصطناعي والمعالجة الحاسوبية للمتلازمات اللفظية والتراكيب الاصطلاحية

د. أيـمن بن أحـمد الغامدي

ملخص الفصل

يُشكل الغموض اللغوي بكافة مستوياته ودرجاته تحديا مستمراً لكثير من مهام المعالجة الآلية للغات الطبيعية؛ ومن هنا بدأ الاهتمام مبكراً بدراسة عدد من الظواهر اللغوية التي تسهم فيه بشكل واضح، ومن أهمها: ظاهرة التراكيب الاصطلاحية والمتلازمات اللفظية، والتي لفتت منذ وقت مبكر انتباه عدد كبير من الباحثين والمهتمين في تخصصات بينية لغوية وحاسوبية متعددة. وفي هذا الفصل نقدم استعراضاً موجزاً لجهود الباحثين في هذا الميدان، من خلال تتبعنا لأهم الدراسات التي اهتمت بالمعالجة الحاسوبية لهذه الظاهرة اللغوية، وسيبدأ الفصل بمقدمة تبين أهمية دراسة هذه الظاهرة وأهم مجالات البحث فيها، ثم يقدم القسم الثاني من هذا الفصل إطاراً نظرياً لدراسة هذه الظاهرة ويشتمل على التعريف العملي، وذكر أهم الخصائص اللغوية المميزة لها في اللغة العربية، بالإضافة إلى استعراض أهم التصنيفات المستعملة للتراكيب الاصطلاحية في مستويات لغوية متعددة. وفي القسم الثالث نقدم استعراضاً لأهم تطبيقات المعالجة الحاسوبية لهذه الظاهرة والتي تلخص المشاكل البحثية الرئيسة التي تتضمن التراكيب الاصطلاحية في أدبيات معالجة اللغات، ويسلط هذا القسم الضوء بشكل خاص على مهمتي الاستخراج والتعرف الآلي، وما يتعلق بالمصادر اللغوية الحاسوبية للتراكيب الاصطلاحية وتطبيقات معالجة اللغات، وأخيراً نختم هذا الفصل بعرض موجز لأبرز التحديات التي لا تزال تشكل عقبة في سبيل الوصول إلى درجات عالية من الدقة في مهام المعالجة الحاسوبية المختلفة لهذه الظاهرة اللغوية المعقدة.

د. أيـمن بن أحـمد الغامدي

أستاذ اللسانيات الحاسوبية المساعد في معهد اللغة العربية في جامعة أم القرى، حاصل على درجة البكالوريوس في اللغة العربية وآدابها، والدبلوم العالي في التربية والتعليم وطرق التدريس في جامعة الطائف، ودرجة الماجستير في اللسانيات التطبيقية في جامعة Essex، ودرجة الدكتوراه في اللسانيات الحاسوبية والذكاء الاصطناعي في قسم الحاسب الآلي في جامعة Leeds في المملكة المتحدة، له عدد من الأبحاث المنشورة في اللسانيات التطبيقية والحاسوبية، ومُنح عدداً من شهادات الشكر والتميز من جهات أكاديمية وخيرية واجتماعية، وله عدد من المشاركات في المؤتمرات العلمية المحلية والدولية، قَدَّم وحَضر العديد من الدورات والورش التدريبية المحلية والدولية في مجالات أكاديمية متنوعة. (https: / /uqu.edu.sa /aamansoori) – (aamansoori@uqu.edu.sa)

١ – المقدمة

تُعد ظاهرة التراكيب الاصطلاحية من الظواهر اللغوية المعقدة التي شغلت كثيراً من الباحثين في عدد من المجالات العلمية المتصلة باللغة، كعلوم اللسانيات المتنوعة مثل: اللسانيات التطبيقية والنفسية، وكذلك في عدد من تخصصات الذكاء الاصطناعي، كاللسانيات الحاسوبية، وعلوم المعالجة الآلية للغات. وتحاول أغلب هذه الأبحاث العمل على تقديم مقترحات علمية وعملية تساعد في تقليل نسبة الغموض اللغوي الذي تسببه هذه الظاهرة في عدد من تطبيقات المعالجة الآلية للغات: كالترجمة الآلية، وتطبيقات التحليل اللغوي المختلفة: مثل التحليل الصرفي والنحوي والدلالي وغيرها.

ويعود الاهتمام بهذه الظاهرة اللغوية لعدد من الأسباب التي من أهمها النسبة الكبيرة التي تشكلها هذه التراكيب في اللغة، وخاصة في اللغة الشائعة التي تستعمل في الحياة اليومية، ففي اللغة الإنجليزية على سبيل المثال تتراوح النسب المقدرة لهذه التراكيب من ٣٠٪ (Biber et al., 1999) إلى أكثر من ٥٠٪ (Erman and Warren, 2000) وفي المعجم الحاسوبي الإنجليزي WordNet شكلت نسبة هذه التراكيب بكل أنواعها حوالي ٤١٪ من إجمالي عدد المداخل المعجمية (Miller et al., 1990).

أما في اللغة العربية فبالرغم من عدم وجود دراسات إحصائية أو نسب محددة لهذه التراكيب كما في الإنجليزية، إلا أن كثرتها تعد كذلك ظاهرة في العربية، وخاصة عند استقراء نتائج البحث في المدونات العربية الضخمة[1]. كما يؤكد أهمية هذه الظاهرة في العربية، العناية المبكرة بها من قبل الباحثين واللغويين العرب، فقديماً ظهر هذا الاهتمام في مؤلفات كثيرة اهتمت بجمع أمثال العرب وحكمهم، وكذلك قصد بعض المؤلفين تفسير ماورد في نصوص الوحيين الكتاب والسنة من الأمثال والحكم، وهذه العناية المبكرة بلا شك تدل على الوعي المبكر عند علماء اللغة بالأهمية البالغة لهذا النوع من التراكيب في فهم اللغة وتفسير معانيها، وكذلك دورها في رفع المستوى اللغوي للكتاب وطلاب العلم، ويوضح جدول ١ عدداً من المصادر العربية القديمة التي اعتنت بهذا النوع من التراكيب.

١– أكدت كثير من الدراسات اللغوية المبنية على مدونات لغوية شيوع هذا النوع من التراكيب في اللغة العربية، ويمكن للمهتم مراجعة الدراسات التالية لمزيد من المعلومات حول هذا الموضوع: (الخولي، ١٩٩٨؛ فايد، ٢٠١٤؛ Alghamdi, 2018 Zaghouani, 2014; Abdou, 2011).

الكتاب	المؤلف	تاريخ وفاته
الأمثال	صحار بن عياش	٦٠هـ
الأمثال	المفضل الضبي	١٧٠هـ
الأمثال	السدوسي	١٩٥هـ
الأمثال	القاسم بن سلام	٢٢٣هـ
الفاخر	ابن سلمة	٢٩١هـ
الأمثال	أبو عكرمة الضبي	٢٥٠هـ
ثمار القلوب في المضاف والمنسوب	أبو منصور الثعالبي	٣٥٠هـ
كتاب أفعل من كذا	أبو علي القالي	٣٥٦هـ
شرح كتاب الأمثال	البكري	٤٨٧هـ
مجمع الأمثال	الميداني	٥١٨هـ
أساس البلاغة	الزمخشري	٥٣٨هـ
ما يعول عليه في المضاف والمضاف إليه	المحبي	١١١١هـ

جدول ١ : أمثلة لكتب جمعت عددا من أنواع التراكيب الاصطلاحية في المصادر العربية القديمة.

وفي الدراسات اللغوية الحديثة ظهر الاهتمام بهذه الظاهرة وما يتصل بها جليا في عدد كبير من الأبحاث التي تناولت هذه الظاهرة من مختلف زواياها، فعلى سبيل المثال، اهتم بعض الباحثين بجمع هذه التراكيب في معاجم خاصة كما في هذه الأمثلة: (أبو سعد، ١٩٨٧؛ إسماعيل وآخرون، ١٩٩٦؛ بشارة،٢٠٠٢؛ أبو داوود،٢٠٠٣؛ حافظ، ٢٠٠٤؛ كامل، ٢٠٠٧، وغيرها)، بينما اتجهت دراسات أخرى إلى تقديم أطر نظرية لدراسة هذا النوع من التراكيب كما في الأعمال التالية:(القاسمي، ١٩٧٩؛ حجازي، ١٩٨٠؛ غزالة، ١٩٩٣؛ هليل، ١٩٩٦؛ ابن عمر، ٢٠٠٧)، وقدمت هذه الأبحاث العديد من المقترحات النظرية فيما يتعلق بتعريف هذه التراكيب اللغوية، وشرح أنواعها، وجمع ما يستعمل من مصطلحات مختلفة في وصفها، بالإضافة إلى دراسة أهم خصائصها وتصنيفاتها في مختلف مستويات التحليل اللغوي.

وقد أكدت بعض الدراسات اللغوية الحديثة المبنية على مدونات ضخمة تمثل اللغة العربية المعاصرة أن التراكيب الاصطلاحية وما يتصل بها من ظواهر لغوية مشابهة

يجب أن تشكل جزءاً أساسياً في كل البرامج اللغوية التي تهدف إلى تحليل النصوص العربية حاسوبيا؛ وذلك لأثرها المهم في تحديد مستويات الغموض اللغوي في مخرجات تطبيقات معالجة اللغة المختلفة(Abdou, 2011; Najar et al., 2015).

وإذا ما أمعنا النظر نجد كذلك أن كثيرا من الكلمات المفردة في اللغة العربية يتوقف فهم معانيها المختلفة على فهم معنى عدد من التراكيب المتصلة بها، وقد تكون بعض معاني هذه التراكيب أكثر شيوعا من معنى الكلمة وهي مفردة، كما يظهر ذلك على سبيل المثال في كلمة «عين»، والتي لا يمكن استيعاب معانيها المتعددة في السياقات اللغوية المختلفة إلا بفهم معاني عدد من المتلازمات اللفظية المتصلة بها، ويمكن توضيح هذه الفكرة من خلال تشبيه الكلمة المفردة «عين» برأس جبل الجليد، والذي قد يظهر في أول وهلة صغيراً ولكنه في الواقع وعند التأمل مجرد قمة لجبل عظيم كما هو موضح في الشكل ١ .

شكل ١ : صورة رمزية لقمة جبل الجليد تظهر فيها كلمة «عين» وأمثلة للتراكيب المرتبطة بها.

ومما يؤكد أهمية دراسة هذه الظاهرة اللغوية وما يتصل بها، ما توصلت إليه عدد من الأبحاث في علم اللغة العصبي والنفسي من نتائج تفيد أن المعجم اللغوي في العقل البشري لا يتكون من مفردات وكلمات معزولة فحسب بل يتمثل في شبكة معقدة من التراكيب والعلاقات المتنوعة التي تمثل المعجم اللغوي للعقل البشري (Wray, 2002 Sinclair, 1991). بالإضافة إلى ذلك، تؤكد كثير من الأبحاث في علم اللغة التطبيقي أن طلاقة متعلم اللغة تعتمد بشكل أساسي على مدى إتقانه ومعرفته بهذه التراكيب وفهم الصلات والعلاقات اللغوية التي تحكمها (Fillmore, 1979; Pawley and Syder,1983; Ohlrogge, 2009).

وفي علوم اللسانيات الحاسوبية ومعالجة اللغات ظهر كذلك اهتمام مبكر بهذه الظاهرة وذلك لدورها المحوري في تحسين الدقة اللغوية لنتائج كثير من تطبيقات المعالجة الآلية للغات، فوجدت في هذا المجال العديد من المشاريع البحثية التي تهدف إلى تضمين هذه التراكيب في مراحل المعالجة الآلية التقليدية للغات[1]، وذلك ببناء معاجم حاسوبية دلالية مختصة بهذه التراكيب كما نجد في الأعمال التالية:(Bar et al., 2014; Constant et al., 2013; Alghamdi, 2018) أو تحسين عمل الخوارزميات الخاصة بالتعرف والاستخراج الآلي لهذه التراكيب من النصوص كما في هذه الأمثلة: (Ramisch 2015; Carpuat and Diab, 2010; Rikters and Bojar, 2017)، وتعود بداية الأبحاث التي وظفت الأدوات الحاسوبية في دراسة هذه الظاهرة اللغوية إلى الستينات الميلادية مع بدايات اختراع الحاسوب وانتشار استعماله، وقد ركزت الأبحاث المبكرة في هذا المجال على تطبيق عدد من الطرق التي تفيد من الحاسوب وقدراته الفائقة في الاستخراج الآلي لعدد من أنواع التراكيب الاصطلاحية بناء على قوالب لغوية محددة مسبقاً كما نجد في الدراسات التالية:(Stevens and Giuliano, 1965; Berry-Rogghe, 1973; Atwell, 1988).

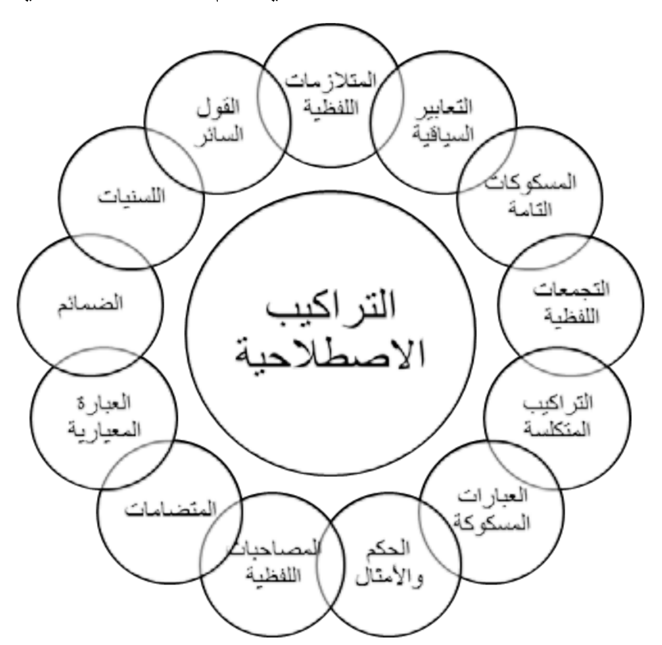
١- تتكون المعالجة الآلية للغات من مجموعة من المراحل المتعارف عليها في اللسانيات الحاسوبية والتي غاليا ما تبدأ بعدد من مهام تحضير النصوص المراد معالجتها ثم توظيف عدد من الخوارزميات في التحليل اللغوي والتي تشمل التقسيم الآلي للكلمات والجمل ثم إرجاع المشتقات الصرفية إلى أصولها ثم تأتي مرحلة الترميز الآلي للوحدات الصرفية بعدد من المعلومات اللغوية المتعلقة بأقسام الكلام والعلاقات النحوية والمعلومات الدلالية، لمزيد من التفاصيل عن مراحل التحليل اللغوي الآلي يمكن مراجعة (حمادة، ٢٠٠٩).

٢- الإطار النظري

في هذا الجزء سنتناول باختصار أهم المقدمات النظرية التي تشكل مدخلاً لفهم هذه الظاهرة اللغوية، وسنقتصر هنا تحديداً على ذكر ما هو مهم لفهم عدد من المشاكل الحاسوبية التي تنشأ عن معالجة هذه الظاهرة في المستويات اللغوية المختلفة.

١ , ٢ تعريف التراكيب الاصطلاحية والمتلازمات اللفظية

من أهم ما يشغل الباحثين في هذه الظاهرة كثرة وتعدد المفاهيم والمصطلحات المستعملة في وصفها، كما يوضح الشكل ٢ أمثلة لعدد من المصطلحات المستعملة لوصف هذه التراكيب والظواهر المتصلة بها في اللغة العربية، ويمكن تبرير هذا التنوع والاختلاف في المصطلح والمفهوم بكثرة الأبحاث وحداثتها في هذه الظاهرة المعقدة، وكذلك شيوع التراكيب الاصطلاحية وتعدد أنواعها، ولذا فكل باحث يحاول أن يقدم تصوراً لهذه الظاهرة اللغوية -متعددة الأوجه- من الزاوية التي يهتم بها أو المشكلة التي يعالجها.



شكل ٢: نماذج من المصطلحات المستعملة لوصف ظاهرة التراكيب الاصطلاحية في اللغة العربية[١].

١- في هذا الفصل يستعمل الباحث مصطلحي (التراكيب الاصطلاحية والمتلازمات اللفظية) ويراد بهما مفهوم واحد وهو الذي نشرحه في هذا الجزء من البحث.

وهنا سنكتفي بسرد عدد من التعريفات المقترحة والتي تتفق في مجموعة من الصفات والخصائص المشتركة للتراكيب اللغوية المستهدفة في سياق هذه الدراسة وهي كما يلي:

- «كلمتين أو مجموعة من الكلمات ترد مع بعضها بعضاً بشكل دائم وثابت في مختلف السياقات»(غزالة، ١٩٩٣ ص.٧).
- «كل عبارة من العبارات المتواترة في اللغة، وقد تكلَّست مكوناتها وتواردت في شكل من أشكال المركبات النحوية المختلفة؛ للدلالة على معنى تعادل قيمته الإخبارية قيمة العلامة اللغوية الواحدة» (ابن عمر، ٢٠٠٧ ص.٤٢).
- «تجمع لفظي (أكثر من وحدة معجمية بسيطة)، يقع في الاستعمال اللغوي باطراد، وله دلالة ثابتة لا تنتج من تجميع دلالات مفرداته المكونة له». (فايد، ٢٠١٤ ص.١١٣).

ومن التعريفات التي يكثر استعمالها خاصة في أدبيات المعالجة الآلية للغات واللسانيات الحاسوبية تعريف Baldwin and Kim (2010: p.269) والذي اعتمد بشكل كبير على تعريف سابق اقترحه Sag et al. (2002) لهذا النوع من التراكيب والذي يمكن ترجمته كما يلي:

«التراكيب الاصطلاحية هي وحدات معجمية ثابتة يمكن تقسيمها إلى وحدات معجمية أبسط منها، وتتميز بظهور المعنى الاصطلاحي أو المجازي فيها والذي يسبب نوعاً من الغموض في أحد مستويات التحليل اللغوي: (المعجمي – التركيبي – الدلالي –الوظيفي– الإحصائي)».

ومن خلال هذه التعريفات وغيرها يمكننا تحديد مجموعة من الخصائص اللغوية التي يمكن استعمالها لتمييز هذا النوع من التراكيب وهو ما سنتناوله في الجزء التالي من هذا الفصل.

٢,٢ الخصائص اللغوية للتراكيب الاصطلاحية

تتميز التراكيب الاصطلاحية التي تشكل هذه الظاهرة اللغوية المعقدة بعدد من الصفات التي تجعلها سببا لمشاكل متعددة في التحليل اللغوي الآلي ومن أهم هذه الصفات ما يلي:

- تعدد مكونات التركيب: فمقتضى كلمة تركيب تعني بالضرورة أنها لابد أن تتكون من وحدتين معجميتين على الأقل، وهذا ما يميزها عن المفردات المنعزلة والكلمات المستقلة، ويرى كثير من اللغويين المحدثين أن التفريق بين مفهومي الكلمة والتركيب في التطبيقات الحاسوبية مثار كثير من الاختلاف والجدل؛ لأن الكلمة يمكن أن يقصد بها تركيب لغوي كامل وخاصة إذا كان المعيار الوحيد للتفريق هو وجود المسافة أو الفراغ الذي يكون بين المفردات، وهو معيار وإن كان سهل التطبيق في مهام المعالجة الآلية للغة إلا أنه غير دقيق خاصة في اللغة العربية التي تتميز بالتداخل الشديد بين الوحدات المعجمية في المفردة الواحدة، كما في تركيب (أرأيتها؟) الذي اجتمعت فيه أربع وحدات معجمية متصلة وليس بينها مسافة في الكتابة، ففي اللغة العربية من الشائع أن نرى جملاً كاملة في صورة مفردة واحدة كما في المثال السابق.

- التواتر وشيوع التلازم بين مكوناتها: من أهم ما تتصف به التراكيب الاصطلاحية أن الوحدات المعجمية المكونة لها غالبا ما تكون متصاحبة في الاستعمال ولو اختلف السياق اللغوي الذي تأتي فيه، وكذلك لا يمكن في أغلب الأحوال استبدال مكوناتها بألفاظ أخرى مرادفة لها.

- المعنى الاصطلاحي: وهذه أهم صفة يمكن بها تمييز هذه التراكيب في مستوى التحليل الدلالي، فما يميز هذه التراكيب أنها تدل على معنى اصطلاحي مختلف عن المعنى الحرفي الذي تدل عليه مكوناتها من الكلمات المفردة. ويُعبر أحيانا عن هذه الصفة بالمعنى الكلي أو الإجمالي للتركيب والذي لا يناسب ولا يتوافق مع دلالة أجزائه، وهذه الصفة توجد في التراكيب بدرجات متفاوتة، فكلما ابتعد المعنى الكلي عن المعاني الحرفية للمفردات، كلما قَل مستوى شفافية التركيب، وينتج عن ذلك مستوى عال من الغموض اللغوي، وخاصة عند

استعمال التحليل الآلي التقليدي المعتمد على معالجة المفردات بشكل مستقل عن المستوى التركيبي لها.

ومثل هذه التراكيب تصعب ترجمتها اعتمادا على ترجمة الكلمات المكونة لها؛ لأنها بمعناها الكلي صارت وحدة معجمية ذات دلالة مستقلة، وقد تنبه اللغويون العرب منذ وقت مبكر جداً لأهمية هذه الخاصية في التراكيب اللغوية، وأثرها البالغ في تحديد المعنى الإجمالي للتركيب فعلى سبيل المثال، ذكر سيبويه في الكتاب عددا من التراكيب اللغوية ثم أكد على ضرورة تلازمها وأثر ذلك على فهم المعنى، فيقول:

«واعلم أنَّ هذه الأشياءَ لا ينفرد منها شيء دون ما بعده، وذلك أنَّه لا يجوز أن تقول: كلّمتُه فاه حتّى تقول إلى فيَّ، لأنَّك إنَّما تريد مشافَهةً، والمُشافهةُ لا تكون إلاّ من اثنين، فإنَّما يَصحّ المعنى إذا قلت إلى فِيَّ، ولا يجوز أن تقول بايعتُه يداً، لأنَّك إنَّما تريد أن تقول: أخَذَ منّي وأعطاني، فإنَّما يَصحَ المعنى إذا قلت: بيدٍ لأنهما عَمَلانِ»ص٣٩٢.

وقد تناولت بعض الأبحاث اللغوية الحديثة كذلك هذه الميزة وأثرها في التحليل الدلالي للتراكيب فقد عبر اللغوي المعروف حسان (١٩٧٣) ص.٣٣١ عن هذا المعنى في التراكيب اللغوية بالتضام والضمائم، فيقول:

«ومن قبيل التضام ما يساق من أمثلة التعبيرات المسكوكة مثل: يضرب أخماسًا في أسداس، ويلقي الحبل على الغارب، ويضع الأمور في نصابها، وغير ذلك من العبارات التي تنوسي فيها ما كان لها من المعنى البياني حتى أصبحت كالأمثال لا تحتمل التغيير، ومن هنا جاء وصفها «بالمسكوكة». وإنما ينبغي ذكر الضمائم هنا؛ لأن الاكتفاء بذكر الكلمة دون ضمائمها لا يصل بالمعجم إلى غايته المنشودة».

٣,٢ تصنيفات وأنواع التراكيب الاصطلاحية

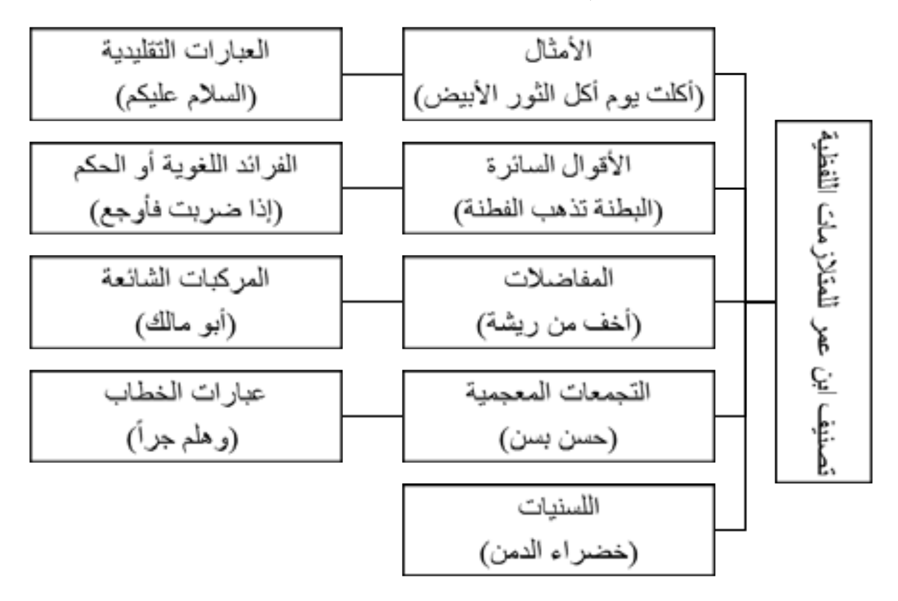
تبعاً لتنوع التراكيب الاصطلاحية وتعدد خصائصها، تنوعت التصنيفات المقترحة لها، وسنتناول في هذا الجزء عددا من التقسيمات المعتبرة وفقا لمعايير ومستويات لغوية مختلفة.

من أشهر التصنيفات استعمالاُ وأكثرها مرونة وسهولة في التطبيق خاصة في مهام المعالجة الآلية للمتلازمات اللفظية، تقسيمها وفقًا لنوع الكلمة الأولى في التركيب أو

تبعاً لما يسمى برأس المركب، وهذا التصنيف يعتمد على اختيار الباحث لأقسام الكلام المعتمدة في بحثه، فعلى سبيل المثال، بناء على التصنيف العربي التقليدي للكلام إلى اسم وفعل وحرف، تكون المركبات تبعاً لذلك: اسمية مثل (ثقيل الدم)، أو فعلية مثل (ركب رأسه)، أو حرفية مثل (على عينك يا تاجر).

واتخذت بعض التصنيفات من عدد الوحدات المعجمية التي تتكون منها معياراً للتقسيم وتبعاً لذلك تكون التراكيب: ثنائية، أو ثلاثية، أو رباعية ... إلخ.

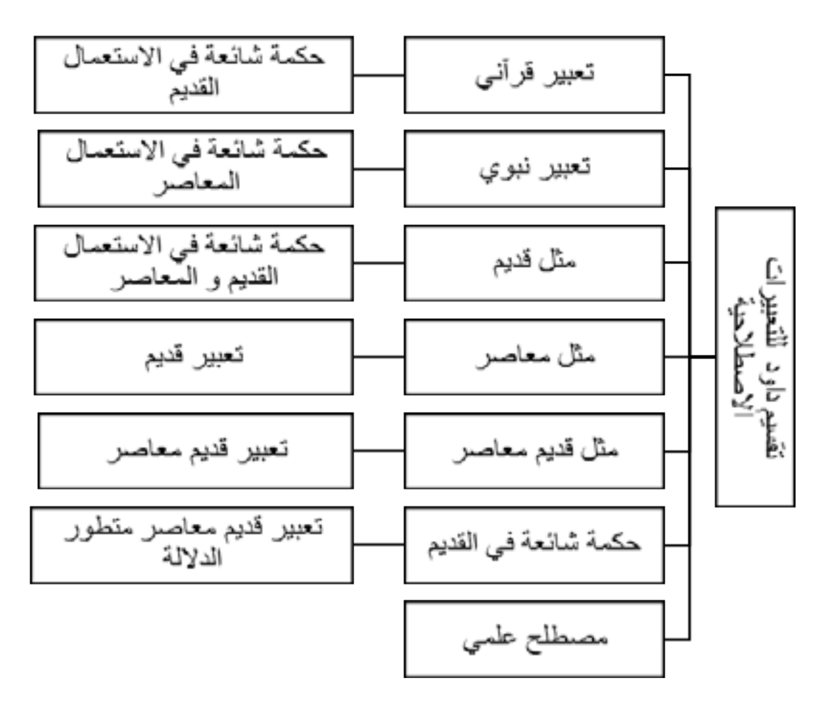
وقدم ابن عمر (٢٠٠٧) تصنيفاً آخر لهذه التراكيب مبني على استقراء عدد كبير من الأمثلة كما يظهر في الشكل رقم ٣ .



الشكل ٣: تصنيف المتلازمات اللفظية وفقاً لابن عمر (٢٠٠٧) ص٢٣-٢٦.

ويظهر في هذا التصنيف التداخل بين بعض هذه الأنواع المقترحة للتراكيب؛ وذلك لعدم وجود معايير لغوية مميزة لها، وهذه هي الحال الغالبة في كثير من التصنيفات المقترحة للمتلازمات اللفظية، وخاصة عندما يكون التفريق بين أنواعها معتمداً على معايير دلالية لا يحصل في الغالب إجماع بين اللغويين في تفسير مفاهيمها، وتبعاً لذلك في تقسيم العبارات بالاستناد عليها.

وفي تصنيف آخر تبنى داود (٢٠١٤) -في معجمه للتعابير الاصطلاحية- تقسيم هذه التراكيب إلى ١٣ نوعاً كما يظهر في الشكل ٤.



الشكل ٤: تصنيف داود للتعبيرات الاصطلاحية في اللغة العربية (داود، ٢٠١٤، ص٢١-٢٢).

ونلاحظ كذلك هنا أن أغلب هذه الأنواع في هذا التصنيف تراعي المستوى الدلالي كمعيار أساسي للتفريق بين الوحدات المعجمية، والمعايير الدلالية غالباً ما تكون مصحوبة ببعض الإشكاليات وخاصة عند محاولة التطبيق العملي كما ذكرنا في التعليق على التصنيف السابق.

ومن التصنيفات المقترحة كذلك تصنيف غريم (٢٠١٤)، والذي يعد من أكثرها تأثراً بالمناهج اللغوية الحديثة، التي يراعي بعضها مدى الاستفادة من هذه التصنيفات في التطبيقات الآلية لمعالجة المتلازمات اللفظية. ويظهر في جدول ٢ تقسيم التراكيب الاصطلاحية إلى أربعة أنواع بناء على معياري الثبات والشفافية[1]، وكما يُلاحظ أن

١- يقصد بثبات التركيب هنا درجة تلازم مكوناته وعدم تغيرها في السياقات اللغوية المختلفة، كما يقصد بالشفافية هنا مستوى استعمال التركيب في معناه الاصطلاحي أو الكلي ومدى بعده عن المعنى الحرفي للكلمات التي يتكون منها.

هناك علاقة عكسية بين هذين المعيارين وذلك يعني أنه كلما كان التركيب أقل ثباتا كان أكثر شفافية، والعكس صحيح.

المعيار	التراكيب الحرة	المتلازمات اللفظية	النحت	التعبير الاصطلاحي
درجة الثبات				
درجة الشفافية				

جدول ٢: تصنيف التراكيب المعجمية بحسب درجة ثباتها وشفافيتها (غريم، ٢٠١٤ ص.٢٩٩).

وكذلك اقترحت غريم تقسيماً آخر للمتلازمات اللفظية مبني على أقسام الكلام لرأس التركيب وكذلك الوظائف النحوية المتنوعة للتعابير الاصطلاحية، ويشمل هذا التصنيف ثلاثة مستويات تنشأ بينها عدد من العلاقات الهرمية كما يظهر ذلك في الجدول رقم ٣.

المستوى الأكبر	المستوى المتوسط	المستوى الأصغر	أمثلة
فعل + اسم	فعل +اسم فاعل		بزغ الفجر
	فعل + مفعول به	فعل + اسم	أسدل الستار
		(فعل +حرف) + اسم	أخذ على عاتقه
		فعل + مفعول مطلق	خضع خضوعاً تاماً
	فعل + حال		تفصد عرقاً
فعل + (حرف + اسم)			استرسل في الحديث
اسم + اسم	اسم + اسم(إضافة)		إطلاق النار
	صفة + اسم معرف (إضافة غير حقيقية)		سليط اللسان
	اسم + صفة	اسم+ نفي+صفة/ اسم	أغلبية ساحقة زيارة غير رسمية
		اسم + نفي + فعل	جزأ لا يتجزأ
اسم + حرف + اسم	اسم + حرف +اسم		صراع على السلطة
	اسم+ من +اسم		عنقود من العنب

جدول ٣:تصنيف المتلازمات وفقاً لأقسام الكلام والوظائف النحوية (غريم،٢٠١٤ ص.٣٠٩-٣١٠).

ومن المفيد أن نشير هنا إلى أن ما يعرف بأسماء الأعلام Named Entity وما يتعلق باستخراجها والتعرف الآلي عليها في النصوص قد أصبح إلى حد ما مجالا علمياً مستقلاً وله أبحاثه ودراساته المتعددة؛ ولذلك فلن يتضمن هذا الفصل معالجة هذا النوع من التراكيب، مع الأخذ في الاعتبار أن نتائج كثير من الأبحاث في هذ المجال تظهر الوصول إلى دقة عالية في المعالجة الحاسوبية لهذا النوع من التراكيب؛ وذلك لتميزها بالثبات اللغوي، وقلة التعقيدات والتغيرات اللغوية التي تطرأ عليها في السياقات المختلفة.

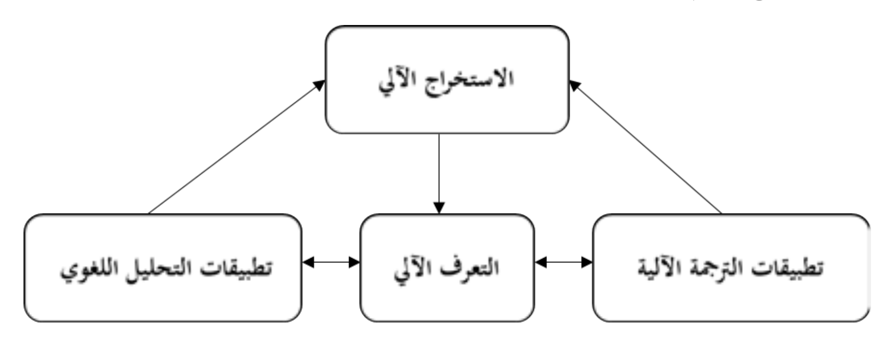
٣- مهام المعالجة الحاسوبية للتراكيب الاصطلاحية

تعددت المناهج المتبعة في تقديم حلول لمشاكل المعالجة الآلية للغات الطبيعية، ووفقا للمنهج المختار، يُحدد الباحث الطرق والأدوات الحاسوبية التي تساعده على إجابة أسئلة البحث وحل مشكلاته وتحدياته، ويرى Dale (2010) أن أغلب الطرق البحثية المتبعة في دراسات معالجة اللغات يمكن تصنيفها لمناهج أربعة رئيسة وهي كما يلي:

- مناهج لغوية تقليدية: تستفيد من المعرفة اللغوية وتقدمها في المعالجة الحاسوبية وتعتمد على محاولة تمثيل قواعد اللغة ونقلها على شكل خوارزميات في تطبيقات المعالجة الآلية للغات.
- مناهج إحصائية: تطبق غالبا على المدونات اللغوية الضخمة، وتستخدم الاختبارات والنماذج والمعلومات الإحصائية لحل مشكلات المعالجة الحاسوبية للغة.
- مناهج معتمدة على تقنيات الذكاء الاصطناعي بالإفادة من تقنيات تعلم الآلة وخوارزميات الشبكات العصبية الاصطناعية وتقنيات التعلم العميق.
- مناهج مختلطة أو متكاملة وهي المناهج التي توظف عددا من المناهج السابقة في الدراسة الواحدة؛ وذلك لتلافي عيوب الاقتصار على منهج واحد في حل المشاكل البحثية المعقدة.

وفي أدبيات المعالجة الآلية للمتلازمات اللفظية من الشائع استعمال هذه المناهج البحثية والإفادة منها في تحسين مهام المعالجة الحاسوبية للتراكيب الاصطلاحية في مختلف

مستويات التحليل اللغوي، ولكي نكون تصوراً شاملا للتطبيقات الحاسوبية في معالجة هذه الظاهرة اللغوية يمكننا أن نشير هنا إلى أهم هذه المهام والعلاقة بينها، فقد ذكر Constant et al. (2017) في مراجعته الشاملة لأدبيات المعالجة الحاسوبية للتراكيب الاصطلاحية أن أغلب الأبحاث في هذا الميدان تسعى لحل مشكلتين أساسيتين وهما: مشكلة الاستخراج الآلي لهذه التراكيب من المدونات اللغوية، وكذلك مشكلة التعرف الآلي عليها في اللغة المكتوبة أو المسموعة، والعمل على حل هاتين المشكلتين في المعالجة الآلية للتراكيب الاصطلاحية يسهم بشكل فعال في تحسين أداء كثير من تطبيقات معالجة اللغات وتعزيز مستوى الدقة في نتائجها. ومن أهم هذه التطبيقات الترجمة الآلية والمهام الحاسوبية الخاصة بالتحليل اللغوي بكافة مستوياته ومراحله، كالتحليل الصرفي والنحوي والدلالي وكذلك ما يتعلق بتقسيم الكلمات وترميزها بالمعلومات اللغوية. ويوضح شكل ٥ المهمتين الرئيستين لمعالجة هذه الظاهرة اللغوية، ونوع العلاقة بينها، حيث يظهر أن العمل على تحسين مهام الاستخراج الآلي يؤدي بالضرورة إلى تحسين نتائج مهام التعرف الآلي على التراكيب الاصطلاحية في النصوص المعالجة.



الشكل ٥: مهام المعالجة الحاسوبية للتراكيب الاصطلاحية والعلاقة بينها (Constant et al., 2017)

وكذلك يشير السهمان مزدوجي الاتجاه إلى أن العلاقة متبادلة بين تطبيقات معالجة اللغة ومهمة التعرف الآلي على التراكيب الاصطلاحية، فتحسن أحد هذه المهام الحاسوبية يؤدي إلى تحسن الأخرى، فعلى سبيل المثال إذا وصلت خوارزمية التعرف الآلي إلى نتائج دقيقة فإن هذا يؤدي إلى تحسن دقة المخرجات اللغوية لتطبيقات التحليل اللغوي والترجمة الآلية، والعكس صحيح كذلك فارتفاع مستوى الدقة في نتائج هذه التطبيقات يؤدي إلى تحسن أداء مهام الاستخراج الآلي للتراكيب الاصطلاحية. ومن

الجدير بالذكر هنا أن نوضح الفرق بين المهمتين الرئيستين لمعالجة التراكيب حاسوبياً؛ وذلك للخلط الذي قد يقع من كثير من الباحثين في هذا المجال، فنقول إنه يمكننا التفريق بينهما بمعرفة الفرق بين مخرجات كل مهمة منهما، فعندما نطبق إحدى تقنيات الاستخراج الآلي فإن المخرجات حينئذ تكون عبارة عن قائمة من التراكيب الـمُستخرجة آليا من مدونة لغوية، وقد تُخزن بعد ذلك في معجم حاسوبي، أو تستعمل كمصدر لغوي في إحدى مهام معالجة اللغة المختلفة.

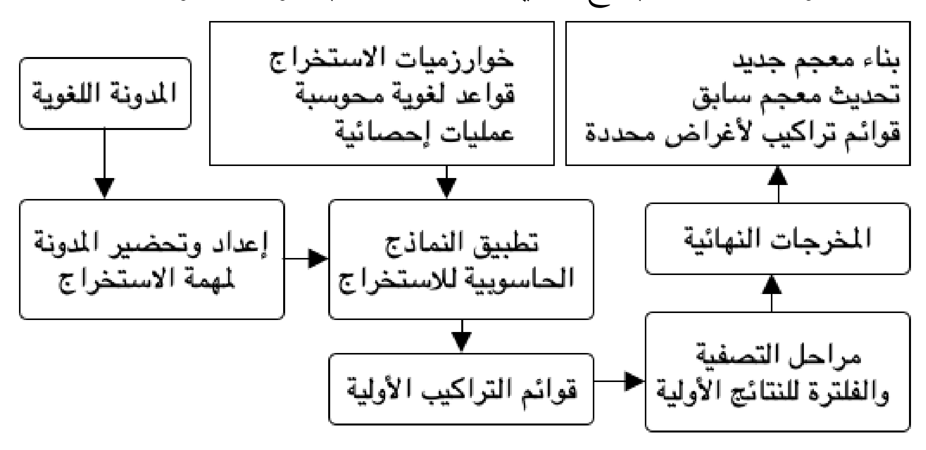
أما فيما يتعلق بمخرجات مهمة التعرف الآلي فهي عبارة عن نصوص موسومة برموز لتمييز التراكيب المتعرَّف عليها، وتختلف هذه الرموز تبعاً لاختلاف آلية التصنيف المعتمدة للتراكيب الاصطلاحية، وكذلك طريقة عمل خوارزميات التعرف الآلي. وسنتناول باختصار في الأجزاء التالية من هذا الفصل عدداً من الأبحاث التي حاولت تقديم إضافة معرفية وتطبيقية في إحدى مهام المعالجة الحاسوبية لهذه الظاهرة اللغوية.

١,٣ مهمة الاستخراج الآلي للتراكيب الاصطلاحية

تعد مهمة الاستخراج الآلي للتراكيب الاصطلاحيةMulti-Word Expressions Extraction من أكثر المهام الحاسوبية بحثا في أدبيات معالجة اللغات المتصلة بالتراكيب الاصطلاحية، ومخرجات هذه المهمة الحاسوبية غالبا ما تكون كما ذكرنا عبارة عن قوائم من التعبيرات الاصطلاحية المستخرجة عن طريق نموذج حاسوبي والتي يمكن بعد ذلك إضافتها لمعجم حاسوبي عام أو استعمالها في بناء معجم خاص بالمتلازمات اللفظية. ومن أهم أهداف هذه المهمة استكشاف التعبيرات الاصطلاحية الجديدة التي تنشأ في الاستعمال اللغوي المعاصر، وكذلك معرفة التطور الدلالي للتعبير الاصطلاحي في سياقات نصية وزمنية مختلفة بالاعتماد على استقراء عدد من المدونات اللغوية.

وإذا ما أردنا أن نضع تعريفاً لهذه المهمة فيمكن أن نقول إنها عبارة عن مجموعة من العمليات الآلية أو غير الآلية والتي يطبق فيها نموذج حاسوبي يتكون من مراحل معالجة متعددة ويتضمن عددا من خوارزميات التنقيب عن البيانات لاستخراج أنواع مختلفة من التراكيب الاصطلاحية المستهدفة في الدراسة بناء على معايير لغوية أو إحصائية محددة.

وقد تعددت المناهج والطرق المستعملة في التطبيق العملي لهذه المهمة، ولكن في الغالب أنها تتفق في ضرورة وجود عدد من الخطوات الأساسية، كالاعتماد على مدونة لغوية في تطبيق نموذج الاستخراج، والأفضل أن تكون المدونة معالجة بأحد برامج التحليل اللغوي الآلي وموسومة برموز لعدد من المعلومات اللغوية الصرفية والنحوية، ثم بعد ذلك يكون تطبيق النموذج الحاسوبي للاستخراج الآلي والذي يتضمن عددا من الخوارزميات والمعادلات الرياضية أو القواعد اللغوية المحوسبة، و يتضمن نموذج الاستخراج كذلك في الغالب عدداً من مراحل المعالجة الآلية وغير الآلية، والتي تتعلق بتصفية وفلترة النتائج الأولية لنموذج الاستخراج الآلي؛ وذلك لاستبعاد مجموعات من أنواع التراكيب غير المناسبة، ويشمل ذلك التراكيب غير المفيدة، أو التي تحوي أخطاء لغوية، أو لا تتناسب مع المعايير المحددة للتراكيب المراد استخراجها.



الشكل ٦: مراحل المعالجة المتبعة في مهمة استخراج التراكيب الاصطلاحية.

يوضح الشكل ٦ الخطوات المتبعة في مهمة الاستخراج الآلي للتراكيب الاصطلاحية، ويظهر في الرسم أن عملية الاستخراج لابد أن تبنى على مدونة لغوية تحوي عددا من النصوص المكتوبة، ومستوى جودة المدونة اللغوية–من ناحية طريقة جمعها، وكذلك مدى تزويدها بالمعلومات اللغوية عن طريق التحليل اللغوي الآلي أو غير الآلي– لها أثر كبير في تحديد مستوى جودة المخرجات النهائية لنموذج الاستخراج.

بعد ذلك تأتي مرحلة تهيئة المدونة وذلك بناء على الهدف المحدد لمهمة الاستخراج، فعلى سبيل المثال، قد يكون الهدف هو الاقتصار على استكشاف التراكيب المرتبطة

بنوع لغوي معين كالتراكيب المستعملة في اللغة العلمية مثلاً، وهنا ينبغي أن يقتصر تطبيق النموذج على هذا النوع من النصوص لتحسين مستوى النتائج المتوقعة لنموذج الاستخراج الآلي.

ثم في مرحلة تطبيق النموذج تُنفذ عدد من العمليات الحاسوبية التي تهدف إلى استكشاف أنواع من التراكيب المقصودة، وتصنيفها بعد ذلك في مجموعات وفقاً لمعايير لغوية أو إحصائية. وفي المرحلة الأخيرة تطبق مجموعة من العمليات الحاسوبية لتصفية النتائج الأولية؛ وذلك باستبعاد العناصر المستخرجة بالخطأ أو بعض أنواع التراكيب غير المرغوب فيها في سياق الدراسة، وتتنوع المخرجات النهائية لمهمة الاستخراج الآلي للتراكيب الاصطلاحية، فقد تكون على شكل قوائم تراكيب مصنفة في فئات متجانسة، أو تكون على شكل مجموعة من الوحدات المعجمية الجديدة التي يمكن إضافتها لمعجم سابق، أو تُستعمل أساساً لمعجم حديث لأنواع محددة من التراكيب الاصطلاحية.

وفي أدبيات معالجة اللغات تعددت وتنوعت الطرق المستعملة في استخراج التراكيب الاصطلاحية، فمن الأبحاث ما يركز على تطبيق طريقة لاستخراج نوع واحد محدد من التراكيب، كالمركبات الاسمية كما في هذه الدراسات: Girju et al., 2005; Salehi et al., 2015 أو المركبات الفعلية كما في الدراسات التالية: Stevenson et al., 2004; Ramisch et al., 2008 ; McCarthy et al., 2003 ومنها ما يتضمن طرقاً هجينة أو متكاملة لاستخراج مجموعة متنوعة من أنواع التراكيب الاصطلاحية كما في هذه الأمثلة: (da Silva et al. 1999; Seretan 2011; Ramisch 2015).

ويمكن تقسيم الطرق المستعملة في استخراج التراكيب الاصطلاحية وفقاً للمنهجية المتبعة في أبحاث معالجة اللغات، والتي سبق ذكرها باختصار في القسم الثالث من هذا الفصل، فبعض الأبحاث تستعمل الطرق التقليدية والتي تركز على ضرورة مراعاة الخصائص اللغوية والمعلومات المعرفية للتراكيب وتعزز من دورها في تطبيق نموذج الاستخراج، وعند تطبيق هذه الطرق من المهم أن تكون المدونة المختارة موسومة بعدد من المعلومات اللغوية التي قد تتضمن أقسام الكلام وأنواع التراكيب والعلاقات النحوية المتعددة، وهذه بعض الأمثلة للدراسات التي استعملت مثل هذه الطرق في استخراج التراكيب الاصطلاحية:(Bartsch, 2004; Cowie, 1998; Mel’ćuk, 1998).

ومن الأبحاث في هذا المجال ما يركز على توظيف المعلومات الإحصائية ويحاول الإفادة منها والتركيز عليها في عملية استكشاف التراكيب اللغوية، وخاصة في ظل توفر مدونات لغوية ضخمة تعزز من دقة المعلومات الإحصائية المستخرجة منها، وهذا المنهج الإحصائي في استخراج التراكيب الاصطلاحية من أكثر المناهج استعمالاً؛ وذلك لسهولة تطبيقه آليا وعدم حاجته إلى التدخل البشري كثيراً في عملية تنفيذ النموذج، وتطبق هذه الطرق بشكل سريع وتؤدي إلى نتائج متميزة لأنها تستثمر القدرات الفائقة للحاسوب في إحصاء ومعالجة كميات ضخمة من البيانات والنصوص اللغوية التي قد تصل إلى بلايين أو تريليونات الكلمات.

ومن أمثلة النماذج المعتمدة على المنهج الإحصائي ما يعرف بنموذج إن-قرام الإحصائي n-gram model وكذلك استعمال ما يعرف بخوارزميات قياس الارتباط والعلاقات الرياضية، والتي لها دور محوري في تحديد مدى تلازم الكلمات في المدونة اللغوية بناء على مجموعة من الاختبارات الإحصائية التي تحدد درجة ارتباطها وقربها في السياقات اللغوية المختلفة، ويوضح جدول ٤ عددا من خوارزميات قياس الارتباط التي يكثر استعمالها مع بيان مراجعها. ومن أهم الدراسات التي اعتمدت على توظيف هذه الخوارزميات ما يلي: (Pecina, 2008; Moirón, 2005; Evert, 2005).

الاسم	المراجع	الخوارزميات
T-score	(Church et al., 1991)	$\frac{f_{xy}-\frac{f_y f_b}{N}}{\sqrt{\int xy}}$
Mutual Information (MI)	(Daille, 1994)	$log_2 \frac{f_{xy}N}{f_x f_y}$
MI3	(Daille, 1994)	$log_2 \frac{\int_{xy}^{3} N}{f_x f_y}$
MI.log_F	(Rychlý, 2008)	$MI - score \times log_{xy}$
logDice	(Rychlý, 2008)	$logDice = 14 + log_2 D = 14 + log_2 \frac{2fxy}{fx+fy}$
Log-likelihood(L.LK)	(Dunning, 1993)	$-2\sum ij \int ij \; log \frac{f_{ij}}{f_{ij}}$

جدول ٤ : أمثلة لعدد من خوارزميات قياس درجة الارتباط مع مراجعها.

ومن الشائع أن تتبع الدراسة التي تهدف إلى تطبيق خوارزميات الارتباط في الاستخراج الآلي للتراكيب الاصطلاحية عددا من الخطوات في مراحل المعالجة المختلفة؛ حتى تضمن الوصول إلى نتائج علمية وأكثر صدقاً في تمثيل بيانات المدونة اللغوية المعتمدة عليها، وهذه الخطوات يمكن تلخيصها في القائمة التالية:

- تحديد أنواع الخوارزميات المستعملة في نموذج الاستخراج.
- تحديد نوع النص الذي ستطبق عليه عملية الاستخراج، فقد تطبق الخوارزميات على النص الأصلي مجردا من أي إضافات، أو قد تطبق على مستوى الرموز اللغوية الصرفية والنحوية المرتبطة بنصوص المدونة.
- تحديد حد أدنى لدرجة شيوع المفردات المستعملة في استخراج التراكيب.
- تحديد حد أدنى لدرجة الارتباط المقبولة وفقا للخوارزميات المستعملة.
- حساب درجة الشيوع العامة للمفردات والتراكيب في المدونة اللغوية.
- تطبيق خوارزميات الارتباط، ثم ترتيب النتائج وتصنيف التراكيب المستخرجة في جداول متعددة بحسب الخوارزمية المستعملة في استكشافها، وكذلك من المفيد أن تفرز التراكيب في قوائم منظمة بطريقة تصاعدية وفقاً لدرجة الارتباط المستخرجة من المدونة اللغوية.
- المقارنة بين نتائج تطبيق خوارزميات الارتباط، وتقييمها من خلال تحديد أفضلها أداءً في مهمة استخراج التراكيب المستهدفة[1].

ومن الطرق الشائعة الأخرى، توظيف عدد من الأنماط والقوالب الصرفية والنحوية للغة في استخراج أنواع مختلفة من التراكيب بناءً على القوالب المعدة مسبقاً، ومن أشهر التراكيب التي يمكن استعمالها كقوالب في استخراج التراكيب الاصطلاحية في اللغة العربية ما يلي: (مضاف +مضاف إليه، فعل+ فاعل، موصوف + صفة) وغيرها الكثير من أنماط الجمل التي قد تأتي فيها المتلازمات اللفظية، ومن الدراسات التي

١- يمكن الرجوع لدراسة (Kyto and Ludeling, 2008) لمزيد من التفاصيل حول هذ الخطوات وطريقة تطبيقها في نماذج الاستخراج الآلي للتراكيب الاصطلاحية.

استعملت هذه الطريقة في استخراج المتلازمات اللفظية: (Castagnoli et al. ,2014; Seretan,2011) وغيرها.

من جهة أخرى، اتجهت بعض الأبحاث في معالجة اللغات إلى توظيف بعض التقنيات المستعلمة في علوم تعلم الآلة والتعلم العميق وما يتصل بها في الاستخراج والتعرف الآلي على التراكيب الاصطلاحية، ومن أمثلة هذه الطرق ما يعرف باستعمال خوارزميات التصنيف التي تعتمد على نماذج التشابه الدلالي Semantic Similarity، وتفيد هذه الطريقة كثيراً في استكشاف التراكيب الاصطلاحية قليلة الشفافية، أو بعبارة أخرى، التراكيب التي غلب على معناها الاستعمال الكلي المجازي الذي لا علاقة له بالمعنى الحرفي للمفردات التي يتكون منها. ويقوم استعمال هذه الخوارزميات على فرضية مفادها أن هناك تشابهاً دلالياً في التمثيل الحاسوبي الدلالي بين التراكيب الاصطلاحية وبعض الكلمات المفردة المرادفة لمعناها الاصطلاحي، فإذا أظهرت نتائج خوارزمية التصنيف تشابهاً دلالياً بين عدد من التراكيب والكلمات المفردة المرادفة لها، فحينئذ يمكن استخراج هذه التراكيب وإضافتها لقوائم التراكيب الاصطلاحية، ويمكن إيضاح هذه المفهوم بمثال للتركيب الاصطلاحي الشائع في اللغة العربية «انتقل إلى رحمة الله»، والذي قد يتشابه دلالياً مع بعض الكلمات المفردة ككلمتي «توفي» أو «مات».

وكذلك قد تستعمل خوارزميات تعلم الآلة المتعلقة بالتشابه الدلالي لتحديد مستوى الشفافية أو مدى الاستعمال الاصطلاحي للتركيب؛ وذلك من خلال مقارنة نتائج التشابه الدلالي بين معاني التركيب الاصطلاحي ومعاني الكلمات المكونة له في سياقات لغوية متفرقة، وكلما كان معنى التركيب بعيدا عن معاني الكلمات المكونة له كان أقل شفافية (Katz and Giesbrecht, 2006).

ومن أهم شروط استعمال هذه الطريقة توفر معاجم حاسوبية موسومة برموز دِلالية للمفردات والتراكيب الاصطلاحية؛ لتتمكن خوارزميات تعلم الآلة من التدرب عليها حتى تصل إلى دقة عالية في مهمة تصنيف العبارات والمفردات إلى مجموعات متشابهة دلالياً، ومن أمثلة الدراسات التي اعتمدت هذه الطريقة: Reddy et al. 2011; Farahmand and Henderson 2016; Riedl and Biemann, 2015 واتجهت مجموعة من الدراسات الأخرى في هذا المجال إلى اعتماد المنهج الهجين أو المتكامل في

استخراج التراكيب الاصطلاحية آليا، والذي يهدف إلى الاستفادة من مميزات عدد من الطرق المختلفة، ويحاول قدر الإمكان التقليل من مشاكل الاعتماد على منهج أو طريقة واحدة، وهذا المنهج من أكثر المناهج استعمالاً في أدبيات معالجة اللغات، وخاصة في ما يتعلق باستكشاف التراكيب الاصطلاحية؛ وذلك لأنها ظاهرة لغوية معقدة ومتشعبة، فمن الأفضل إذا أردنا الوصول إلى نتائج أكثر دقة في المعالجة الحاسوبية لها، أن نُوظف في تصميم نماذج الاستخراج كل التقنيات والطرق المتاحة. وكذلك من فوائد اعتماد هذا المنهج أنه يساعد على مراعاة الخصائص اللغوية للتراكيب المختلفة؛ وذلك بتخصيص كل نوع من التراكيب بطريقة معينة تكون هي الأنسب لخصائصه والأكثر فائدة في معالجته الحاسوبية.

ومن الأمثلة على دراسات استخراج التراكيب الاصطلاحية في اللغة العربية، دراسة(2010) Attia et al. والتي طبقت فيها ثلاث طرق لاستخراج التراكيب الاصطلاحية آلياً بالاعتماد على عدد من التقنيات الإحصائية واللغوية. الطريقة الأولى في هذه الدراسة كانت متأثرة بدراسة (2009) Zarrieß and Kuhn وكانت تستهدف استخراج التراكيب الاصطلاحية قليلة الشفافية، وذلك من خلال ترجمة عناوين موسوعة ويكيبيديا العربية إلى عدد من اللغات الأجنبية وبعد ذلك وبناء على نتائج الترجمة يُصنف العنوان -الذي تكون ترجمته كلمة مفردة في إحدى اللغات المقابلة- عبارة اصطلاحية؛ وذلك بناء على الفرضية التي تدعي أن التركيب الاصطلاحي تكون ترجمته في اللغات الأخرى غالباً «كلمة مفردة»، ويوضح جدول ٥ عدداً من التراكيب المستخرجة باستعمال هذه الطريقة مع ترجمتها إلى الإنجليزية.

العبارة الاصطلاحية	الترجمة
فقر الدم	Anaemia
التهاب القولون	Colitis
ورق الحائط	Wallpaper
قمرة القيادة	Cockpit
فريق عمل	Teamwork

جدول ٥: نماذج من العبارات المستخرجة باستعمال طريقة الترجمة والمقارنة إلى اللغات الأجنبية.

أما الطريقة الثانية، فتبعاً لدراسة Vintar and Fiser (2008) استعمل الباحث الترجمة الثنائية بين العربية والانجليزية كوسيلة لاستكشاف تراكيب اصطلاحية جديدة في اللغة العربية، فبناء على افتراض أن العبارة الاصطلاحية في لغة ما قد تكون كذلك عندما تترجم إلى لغة أخرى، ترجمت الدراسة التراكيب الاصطلاحية المصنفة في المعجم الحاسوبي الدلالي شبكة الكلمات (PWN) Princeton WordNet2[1] إلى اللغة العربية، وبهذه الطريقة تمكن الباحث من استخراج أكثر من ١٣ ألف عبارة اصطلاحية. وفي الطريقة الثالثة استعمل الباحث المنهج الإحصائي من خلال الاعتماد على تطبيق نموذج استخراج آلي يوظف عددا من خوارزميات قياس درجة الارتباط، وقد اعتمدت الدراسة على استخراج التراكيب الاصطلاحية بهذه الطريقة من مدونة عربية تتكون من أكثر من ٨٤٨ مليون كلمة[٢].

وفي دراسة أخرى اقترح AlSabbagh et al., (2014) طريقة لاستخراج عدد من التراكيب الفعلية في اللغة العربية والتي تماثل معاني الأفعال الناقصة في الإنجليزية modal verbs، وقد طبق البحث المنهج الإحصائي الآلي في استخراج التراكيب المستهدفة، وطبق نموذج الاستخراج على عدد من المدونات العربية يصل عدد كلماتها إلى أكثر من ٣٥ مليون كلمة. وقدم Alghamdi and Atwell (2016) مقارنة لتقييم استعمال عدد من خوارزميات الارتباط في استخراج المتلازمات اللفظية في اللغة العربية، وكذلك قامت الدراسة بقياس تأثير عامل شيوع الكلمات المستعملة على أداء خوارزميات الارتباط، وأظهرت نتائج هذه الدراسة أن خوارزمية MI.log_f للارتباط كانت الأفضل أداء في الاستخراج الآلي للمتلازمات اللفظية، وكذلك بينت النتائج تحسناً ملحوظاً لأداء هذه الطريقة عندما تنفذ على مستوى الكلمات الأكثر شيوعاً في المدونة اللغوية.

ولاتزال اللغة العربية في حاجة ملحة إلى مزيد من الدراسات التطبيقية في هذا المجال، تُوظف فيها أحدث الطرق الحاسوبية لاستخراج التراكيب الاصطلاحية من المدونات اللغوية الضخمة؛ وذلك لبناء مصادر لغوية شاملة تعزز من أداء مهام

١- رابط المعجم: http: / /wordnet.princeton.edu.

2- http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp? catalogId=LDC2009T30

معالجة اللغات آلياً. وأخيرا، نؤكد هنا أن ما قُدِّم في هذا الجزء من استعراض سريع لمهمة استخراج التراكيب الاصطلاحية ومراجعة لبعض الدراسات المتصلة بها ما هو إلا نبذة مختصرة عن مجال بحثي واسع تعددت فيه الأبحاث وتداخلت مع عدد من العلوم اللغوية والحاسوبية، كعلم الدلالة وتحليل الخطاب وعلوم التنقيب عن البيانات والتحليل الآلي للنصوص وغيرها كما هي طبيعة أغلب الدراسات البينية في اللسانيات الحاسوبية.

٢, ٣ مهمة التعرف الآلي على التراكيب الاصطلاحية

تُعَدُّ مهمة التعرف الآلي على التراكيب الاصطلاحية MWE Identification جزءاً مهماً من أغلب تطبيقات معالجة اللغات؛ وذلك لما تُشكله من أهمية قصوى في تحسين مستوى جودة ودقة المخرجات النهائية لمهام التحليل اللغوي الآلي المختلفة، فعلى سبيل المثال، في كثير من أنظمة معالجة اللغات والترجمة الآلية غالباً ما تتضمن المراحل الأولية لمعالجة النص نوعاً من أنواع التعرف الآلي على التراكيب الاصطلاحية؛ وذلك لأهمية تخصيص هذا النوع من التراكيب بمعالجة حاسوبية خاصة تقلل من تأثيرها على درجة الغموض اللغوي في المخرجات النهائية لهذه الأنظمة.

وكما أوضحنا في القسم الثالث من هذا الفصل وخاصة من خلال الشكل رقم ٥ أن هناك تداخلاً وصلة دائمة بين مهمتي الاستخراج والتعرف الآلي، فنتائج الاستخراج الآلي تُسهم في تحسين عمل تطبيقات التعرف الآلي والعكس صحيح، وللتفريق بينهما يمكن القول أن مخرجات التعرف الآلي غالباً ما تكون عبارة عن إضافة مجموعة من الرموز الخاصة للتراكيب الاصطلاحية المتعارف عليها في النص المعالج حاسوبياً، ولذلك تعد تطبيقات التعرف الآلي نوعا من أدوات الترميز اللغوية الآلية الخاصة بهذا النوع من التراكيب؛ لأن مهمتها تركز على تصميم خوارزميات لإضافة علامات يمكن تمييزها عن باقي التراكيب اللغوية.

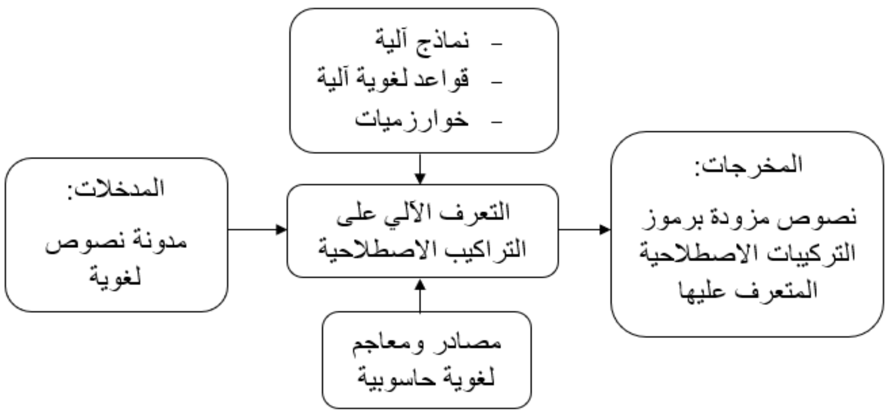
ويوضح شكل ٧ نموذجاً مفترضاً لأحد مخرجات برامج التعرف الآلي حيث يظهر النص موسوماً بعلامات للتراكيب الاصطلاحية. ومن المتعارف عليه في هذا المجال أن برامج الترميز اللغوي الآلي تكون مخرجاتها على قسمين: الترميز النصي والترميز المستقل، ففي النوع الأول تكون الرموز مصاحبة للنص الأصلي، أما النوع الثاني فتكون

الرموز فيه مستقلة في ملفات خاصة بها ومصحوبة بأرقام تشير إلى مواضع هذه الرموز في النصوص الأصلية، ولكل نوع أماكن يحسن استعماله فيها بحسب التطبيق المستهدف من قبل المرمز الآلي.

من المؤكد أنه لن	_يُلقي له بالاً_	وهو يعلم أنّ شخصاً كهذا	_أكل عليه الدهر وشرب_
	MWE		MWE
والضرب فيه حرام	وربما اعتبره	_شاة لا يضرها السلخ بعد مماتها_	
MWE		MWE	

الشكل ٧: مثال لأحد مخرجات التعرف الآلي على التراكيب الاصطلاحية باستعمال الترميز النصي.

ويعتبر توفر المعاجم الحاسوبية للتراكيب الاصطلاحية عاملاً محورياً في تحسين أداء النماذج الحاسوبية المصممة للتعرف الآلي، حيث يمكن بناء خوارزمية تتعرف بسهولة على التراكيب الجديدة المماثلة للتراكيب الاصطلاحية المخزنة في معجم مُعد مسبقاً عن طريق البحث في النص المعالج عن تراكيب مماثلة لمدخلات المعجم الحاسوبي المستعمل في مهمة التعرف الآلي، ويوضح شكل ٨ أهم المراحل التي تتكون منها برامج التعرف الآلي على التراكيب الاصطلاحية، حيث يمكن ملاحظة أن مهمة التعرف الآلي تشتمل على عدد من عمليات معالجة حاسوبية وتستفيد كذلك من المصادر اللغوية المتنوعة ذات الصلة بالتراكيب المراد التعرف عليها.



شكل ٨: المكونات الأساسية لمهمة التعرف الآلي على التراكيب الاصطلاحية.

و نؤكد هنا ما ذُكر في شرح الطرق المستعملة في الاستخراج الآلي للتراكيب، أن الدراسات في هذا المجال استفادت كذلك من كل التقنيات ومناهج البحث المستعملة في مهام معالجة اللغات، ووفقاً لذلك تنوعت المناهج المعتمدة في التطبيقات المصممة للتعرف الآلي، فمنها ما يعتمد الطرق التقليدية القائمة على كتابة قواعد لغوية آلية للتعرف على بعض أنواع التراكيب، ومنها ما يوظف قدرات الحاسوب الفائقة في البحث والمقارنة فيعتمد على المعاجم الحاسوبية المعدة مسبقاً في التعرف الآلي، ومنها ما يوظف مجموعة من خوارزميات تعلم الآلة أو التعلم العميق في تحسين مهمة التعرف الآلي وتوسيع نطاق التراكيب التي يمكن التعرف عليها دون إشراف أو استعانة بمصادر لغوية معدة مسبقاً.

من أقدم الطرق استعمالا في هذا المجال ما يعرف بالطرق المعتمدة على القواعد اللغوية المحوسبة حيث يستفاد فيها من الخصائص اللغوية للتراكيب المستهدفة في بناء قواعد لغوية آلية تمكن البرنامج من التعرف على التراكيب الموافقة للقاعدة المبرمجة مسبقاً، وتتضمن في الغالب برامج التعرف المعتمدة على القواعد المراحل المعتادة في المعالجة الآلية للنص، كتقسيم الكلمات إلى أصغر وحدات صرفية وإرجاع المشتقات إلى أصولها وتزويد النص بالرموز الخاصة بأقسام الكلام والعلاقات النحوية بناء على نتائج التحليل الآلي، وبعد ذلك يكون تطبيق خوارزميات التعرف المعتمدة على مقارنة النص المعالج بقوالب القواعد المخزنة في البرنامج، ومن أهم الدراسات التي تأثرت بهذه الطرق في التعرف الآلي على التراكيب اللغوية دراسةGhoneim and Diab (2013) التي وظفت عددا من تقنيات التعرف على التراكيب في اللغة الإنجليزية والعربية لتحسين نتائج نظام إحصائي للترجمة الآلية بين اللغتين وقد أظهرت نتائج هذه الدراسة تطوراً ملحوظاً عند المقارنة بين نتائج الترجمة قبل وبعد دمج التراكيب الاصطلاحية المتعرف عليها في نظام المترجم الآلي. لكن من المهم هنا التنبيه على أن من أبرز عيوب هذه الطرق صعوبة تعاملها مع التراكيب المتغيرة صرفياً أو نحوياً، وكذلك صعوبة الاستفادة منها في معالجة التراكيب غير المتصلة والتي قد تتنوع فيها الكلمات الفاصلة بين أجزاءها، وهذا النوع من التراكيب لا يمكن التعرف عليه آليا بمجرد استعمال خوارزميات البحث والمطابقة أو التقنيات المعتمدة على قواعد لغوية ثابتة.

ومن الطرق الأخرى المعتمدة على التصنيف المبني على تعلم الآلة في التعرف الآلي، طريقة تمييز المعاني المختلفة للتراكيب Sense Disambiguation Method ، والتي تستعمل فيها خوارزمية التعرف الآلي عددا من التقنيات الإحصائية لاستخراج مجموعة من المعلومات الدلالية عن استعمال التركيب في سياقات لغوية مختلفة، ومن خلال هذه المعلومات تُصنف التراكيب المستهدفة في المعالجة إلى عدة مجموعات بناء على المعلومات الإحصائية عن استعمالاتها المختلفة، ومن ثم تظهر في النتائج التراكيب الاصطلاحية في مجموعات مستقلة متشابهة دلالياً وفقاً لمعلومات وسياق استعمالها، ويتم التركيز في هذه الطرق غالباً على التعرف على التراكيب قليلة الشفافية أو بعبارة أخرى التراكيب المستعملة غالباً في معانيها المجازية. فعلى سبيل المثال قدم Hashimoto and (2008) Kawahara مقترحاً لنظام آلي مبني على عدد من خوارزميات تعلم الآلة يمكنه التعرف الآلي والتفريق بين الاستعمالات الحقيقية والمجازية لعدد من التراكيب الاصطلاحية في اللغة اليابانية. وتتطلب هذه الطرق كمثيلاتها وجود معاجم حاسوبية أو مدونات لغوية موسومة بالمعلومات اللغوية وخاصة ما يتعلق بمعانيها الدلالية في سياقات مختلفة، ليمكن من خلالها تدريب خوارزميات تعلم الآلة على التمييز بين معاني التراكيب في السياقات اللغوية المتعددة.

كذلك توظف بعض الطرق المستعملة في التعرف الآلي على التراكيب الاصطلاحية معلومات التحليل الصرفي والنحوي الآلي في تعزيز دقة الخوارزميات المصممة لهذه المهمة، ومن الأمثلة على ذلك دراسة .Green et al (2013) التي طبقت نموذجاً للتحليل اللغوي الآلي يتضمن الاستفادة من المعلومات اللغوية الصرفية والنحوية في تحسين مستوى الدقة في التعرف الآلي على التراكيب الاصطلاحية في اللغة العربية والفرنسية، وقد وجدت الدراسة كذلك في التجربة المطبقة على عينة من النصوص اللغوية أن نتائج التحليل اللغوي كذلك تأثرت إيجابياً عند دمج التعرف الآلي على التراكيب في مراحل برنامج التحليل اللغوي الآلي المختلفة. وقد تعددت الطرق المقترحة لتقييم النماذج والبرامج الحاسوبية لمهام الاستخراج والتعرف الآلي، وذلك تبعاً لتعدد الطرق المستعملة في هذه التطبيقات، ومن أشهر طرق التقييم في مهمة الاستخراج الآلي ما يلي:

- التصنيف اليدوي للنتائج من قبل الخبراء واللغويين المختصين.
- المقارنة الآلية بمعاجم حاسوبية معدة مسبقا، ومن أبرز عيوب هذه الطريقة صعوبة توفر معاجم محوسبة شاملة للتراكيب الاصطلاحية مما يقلل من فعاليتها ومستوى تغطيتها في تقييم المخرجات الصحيحة التي قد لا توجد في المعاجم المستعملة.
- قوائم التراكيب الاصطلاحية المصممة لمهمة استخراج محددة، وهذه الطريقة مفيدة جدا عندما يكون الهدف هو قياس مدى فعالية نموذج استخراج محدد في سياق لغوي خاص، كاستخراج عدد من التراكيب التي تستعمل بكثرة في الكتابات العلمية والرسائل الأكاديمية على سبيل المثال.
- التقييم المتكامل أو الهجين، والذي يوظف عدداً من الطرق السابقة في تقييم النموذج الحاسوبي، وغالبا ما تستعمل هذه الطريقة عند تعذر الاعتماد على طريقة واحدة لأسباب عملية متعددة.

ويكثر استعمال طرق التقييم المعتادة في نماذج التنقيب عن البيانات لتقييم مخرجات التعرف الآلي على التراكيب الاصطلاحية، حيث تتم مقارنة نتائج المرمز الآلي للتراكيب الاصطلاحية بنصوص موسومة برموز للتراكيب الاصطلاحية من قبل الخبراء، أو بالاعتماد على مصادر لغوية مخصصة للتراكيب الاصطلاحية، وتستعمل في هذا التقييم غالباً درجات القياس المعروفة كدرجة الدقة والاستدعاء ودرجة إف (Precision, Recall, and F-measure).

كذلك تستفيد بعض الدراسات من استعمال ما يعرف بالتقييم التطبيقي، والذي يعتمد على تقييم النموذج الحاسوبي من خلال قياس مدى تأثيره في جودة مخرجات أحد تطبيقات معالجة اللغة ذات الصلة بالتراكيب الاصطلاحية، كالتحليل اللغوي أو الترجمة الآلية، ومن ثم يكون الحكم على مدى فعالية النموذج ودقة نتائجه بناء على مدى تأثيره الإيجابي في تطبيقات معالجة اللغات.

ومن الطرق المستعملة مؤخراً في تقييم وتحسين أداء تطبيقات الاستخراج والتعرف الآلي ما يعرف بطريقة المهمة المشتركة Shared Task، والتي غالباً ما تكون جزءًا من

الفعاليات العلمية الملحقة بالمؤتمرات الأكاديمية المختصة في اللسانيات الحاسوبية وعلوم المعالجة الآلية للغات، فيقوم المنظمون للمؤتمر بعرض مهمة محددة للمختصين المشاركين في المؤتمر كالاستخراج الآلي للتراكيب الاصطلاحية الاسمية على سبيل المثال، ويُطلب بعد ذلك من المشاركين تنفيذ هذ المهمة باستعمال إجراءات وخطوات مشتركة تُشرح لهم بالتزامن مع دعوات المشاركة في المؤتمر، وعند الانتهاء من تنفيذ هذه المهمة، يتدارس المشاركون نتائج تطبيقاتهم المختلفة، ومن خلال المقارنة بين النتائج يتوصل المشاركون إلى تقييم وتصور لأداء تطبيقاتهم في سياقات لغوية متعددة، وقد تتعدد كذلك اللغات المستعملة في المهمة الواحدة، فيكون التقييم حينئذ مفيداً لمعرفة ما إذا كان بالإمكان تعميم النتائج لتشمل لغات أخرى.

٣,٣ التراكيب الاصطلاحية والمعاجم الحاسوبية

لتوفر المصادر اللغوية الآلية دورٌ أساسي في تحسين كثير من نتائج تطبيقات معالجة اللغات، وخاصة ما يتعلق منها بمعالجة التراكيب الاصطلاحية، فكثير من التطبيقات التي سبق الحديث عنها في هذا الفصل تعتمد دقة النتائج فيها على مدى توفر هذه المصادر اللغوية ومستوى جودتها؛ ولهذه الأهمية نجد في أدبيات المعالجة الحاسوبية للتراكيب الاصطلاحية عدداً كبيراً من الأبحاث التي تهتم ببناء مصادر لغوية حاسوبية للتراكيب الاصطلاحية، وتتضمن كذلك في بعض الحالات أنظمة تمثيلية للمعلومات اللغوية التي تضاف لها في المستويات الصرفية والنحوية والدلالية وغيرها؛ وذلك لتوسيع نطاق الاستفادة من هذه المصادر في تطبيقات معالجة اللغة المتعددة.

وقد قَدَّم Losnegaard et al. (2016) مراجعة شاملة للمصادر الحاسوبية المتوفرة للتركيب الاصطلاحية، واعتمد الباحثون في معرفة هذه المصادر على قواعد بيانات المصادر اللغوية على شبكة الإنترنت[١]، وكذلك باستعمال استبانة خاصة[٢] صُممت لهذا الغرض ووزعت على الباحثين والمهتمين في قوائم بريدية متنوعة، وفي

١- من أهم المصادر التي رجعت لها لدراسة قواعد البيانات التالية:

- META-SHARE: the ILSP managing node.
- ELRA: European Language Resources Association.
- SIGLEX-MWE: the MWE community website.

٢- الاستبانة متاحة على الإنترنت ويمكن الرجوع لها من خلال هذا الرابط: https://goo.gl/eYz8qL

نتائج هذه المراجعة يُلاحظ وجود تنوع في هذه المصادر؛ وذلك وفقاً للأغراض التي أنشئت من أجلها، فمنها على سبيل المثال، ما يكون على شكل وحدات معجمية مجردة من سياقاتها اللغوية، ومنها ما يتضمن جملاً طويلة لشرح معاني الدلالات المختلفة للتركيب في سياقات لغوية مختلفة، ومن هذه المصادر كذلك ما يقتصر على لغة واحدة ومنها ما يشمل لغات متعددة. وتُوفر بعض هذه المصادر معلومات لغوية إضافية عن التركيب أو الوحدات المعجمية بالاعتماد على نظام حاسوبي لتمثيل البيانات اللغوية، وقد تعتمد بعض هذه المصادر على الأنظمة القياسية لتمثيل المصادر اللغوية كالنظام القياسي المعروف[١] Lexical Mark-up Framework، ومن أهم فوائد اعتماد هذه الأنظمة القياسية في تمثيل البيانات اللغوية، سهولة استعمال المصدر اللغوي وتوظيفه في تطبيقات حاسوبية مختلفة دون الحاجة لإضافة الكثير من التغييرات على المصادر الأصلية.

وقد تعددت المعاجم الحاسوبية المطورة في اللغة العربية، فعلى سبيل المثال قدم Alghamdi and Atwell (2016) نظاما لتمثيل البيانات اللغوية للتراكيب الاصطلاحية في اللغة العربية وذلك للوصول إلى تمثيل حاسوبي شامل لهذه الظاهرة اللغوية يراعي الخصائص الفريدة للغة العربية بمختلف مظاهرها ومستوياتها اللغوية، وفي دراسة أخرى طَوَّر Najar et al. (2016) معجماً للمركبات الاسمية في اللغة العربية مع تمثيل لها في البيئة الحاسوبية الخاصة بمهام معالجة اللغات والمعروفة بنوج[٢] Nooj، وقد اتجهت عدد من الدراسات الأخرى في هذا المجال إلى نشر قوائم لعدد من التراكيب التي تم التعرف عليها باستخدام طرق الاستخراج الآلي المتنوعة في مواقع خاصة أو ضمن قواعد البيانات اللغوية على الإنترنت، كما في هذه الأمثلة: (Hawwari et al.,2014; Attia, 2006; Abdu, 2011). وعلى الرغم من تعدد الدراسات في هذ المجال إلا أن اللغة العربية لاتزال في حاجة إلى المزيد من البحث والجهود العلمية المؤسسية لبناء مصادر لغوية حاسوبية حديثة تمثل اللغة العربية بكافة مستوياتها كما

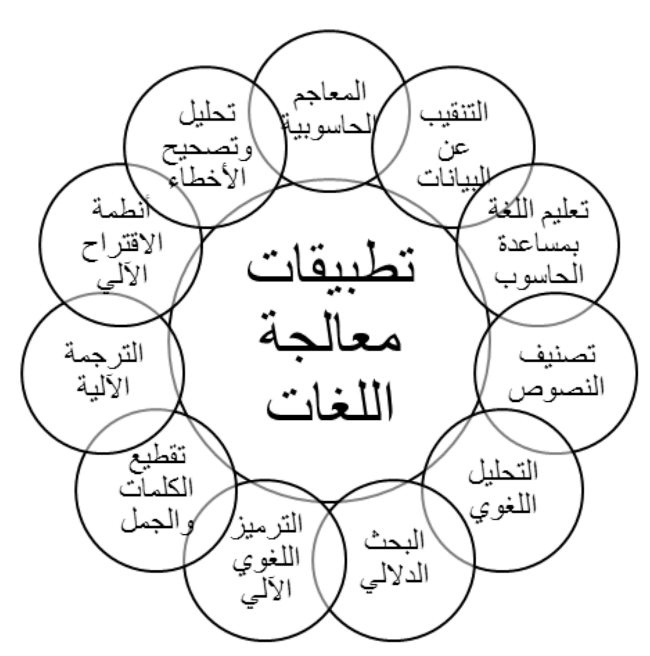
١- لمزيد من المعلومات حول هذ النظام وطريقة تطبيقه على لغات متعددة ومنها العربية يمكن الرجوع إلى كتابه الأساسي (Francopoulo, 2013)

٢- تقدم هذه البيئة مجموعة من الأدوات الحاسوبية لمعالجة اللغة بكافة مستوياتها، ولمزيد من التفاصيل يمكن الرجوع لهذا المصدر (Silberztein, 2016)

توفر معلومات دقيقة عن التطور الدلالي لاستعمال التراكيب الاصطلاحية في الأزمنة والأماكن المختلفة التي تستعمل فيها اللغة العربية، وتقدم كذلك أنظمة تفصيلية لتمثيل البيانات اللغوية.

٤, ٣ التراكيب الاصطلاحية وتطبيقات معالجة اللغات

أثبتت كثير من الأبحاث والتجارب التطبيقية في كثير من أدبيات معالجة اللغات واللسانيات الحاسوبية (Tan and Pal, 2014; Monti, 2015; Carpuat and Diab,2010) أن المعالجة الحاسوبية الكافية لهذه الظاهرة اللغوية لها أثر إيجابي كبير في تحسين مخرجات كثير من مهام معالجة اللغات آلياً؛ وذلك لدورها المحوري في تقليل نسبة الغموض اللغوي في النتائج الأخيرة لهذه التطبيقات المختلفة، والتي من أهمها تطبيقات التحليل اللغوي بمستوياته المختلفة وكذلك الترجمة الآلية، ويوضح الشكل رقم ٩ أمثلة لعدد من التطبيقات التي يمكن فيها دمج المعالجة الحاسوبية لهذه الظاهرة اللغوية.



الشكل ٩: أمثلة لتطبيقات المعالجة الآلية للغات.

فعلى سبيل المثال، أثبتت دراسة Ghoneim and Diab (2013) تحسنا ملحوظاً في مخرجات نظام الترجمة الآلية بين اللغة العربية والإنجليزية عند دمج معالجة التراكيب الاصطلاحية في نموذج الترجمة الإحصائي، و طبقت الدراسة أربع تقنيات لدمج التراكيب الاصطلاحية في نظام الترجمة الآلية وذلك وفقاً للخصائص اللغوية للتراكيب الاصطلاحية المستهدف إدماجها في برنامج الترجمة. وأوضحت النتائج تأثيراً إيجابياً لدمج التراكيب الاصطلاحية في تحسن جودة نتائج تطبيق الترجمة.

وفي أدبيات الترجمة الآلية، تعددت الطرق المطبقة في دمج معالجة التراكيب الاصطلاحية في أنظمة الترجمة الآلية المتعددة، فمنها ما يعتمد على التعرف الآلي الأولي لهذه التراكيب قبل بداية الترجمة، ومنها ما يعتمد على استخراج هذه التراكيب بعد عملية الترجمة، واتجهت دراسات أخرى إلى دمج معالجة التراكيب الاصطلاحية في داخل نظام الترجمة؛ بتفعيل خوارزميات التعرف الآلي على هذه التراكيب باعتبارها إحدى مراحل النموذج الخاص بالترجمة الآلية.

وتُعد برامج التحليل اللغوي الآلي من أكثر تطبيقات معالجة اللغة إفادة من دمج معالجة التراكيب الاصطلاحية؛ وذلك للدور المهم لهذه المعالجة في تحسين نتائج التحليل الصرفي والنحوي والدلالي، والعكس صحيح فتحسن جودة مهام التحليل اللغوي تؤدي إلى تحسن مهام المعالجة الآلية الرئيسة لهذه التراكيب كالاستكشاف والتعرف الآلي. وقد تعددت كذلك الأساليب المطبقة لدمج التراكيب الاصطلاحية في نموذج التحليل اللغوي الآلي فمنها ما يعتمد على الاستخراج الآلي لهذه التراكيب ووضعها في قوائم خاصة بعد نهاية عمليات التحليل اللغوي المتنوعة وذلك لضمان عدم تأثرها بالمعالجة الآلية المعتادة للمفردات والتراكيب في اللغة، ومنها ما يُوظف عدداً من تقنيات التعرف الآلي على هذه التراكيب قبل أو بعد أو في أثناء تطبيق نموذج التحليل اللغوي. وفي اللغة العربية أثبتت دراسة Attia (2006) تحسنا ملحوظاً في التحليل الآلي اللغوي للغة العربية عند دمج المعالجة الحاسوبية لبعض أنواع التراكيب الاصطلاحية، وأوصت الدراسة بتعدد أساليب هذ الدمج لتشمل كافة أجزاء نموذج التحليل اللغوي بداية بمرحلة إعداد وتحضير النص والتقسيم الآلي للكلمات والجمل وانتهاء بالمراحل المتقدمة كالتحليل النحوي والدلالي والوظيفي للنصوص.

٤- عقبات وتحديات

على الرغم من التقدم الذي يمكن ملاحظته في المعالجة الحاسوبية للتراكيب الاصطلاحية إلا أن البحث في هذ المجال لايزال يواجه عدداً من التحديات والمشكلات المعقدة المفتوحة التي تتطلب جهوداً وحلولاً علمية وعملية لتحسين معالجة هذه الظاهرة في تجلياتها وأنواعها المختلفة، وفي هذا القسم سنشير باختصار إلى أهم هذه التحديات.

من أهم التحديات البحثية في مهام الاستخراج الآلي للتراكيب الاصطلاحية، أنه على الرغم من الفائدة الكبيرة التي قدمتها عدد من خوارزميات الاستخراج المتنوعة وخاصة التي تشكل جزءاً من النماذج الآلية المعتمدة على المعلومات الإحصائية كمعادلات الارتباط ونموذج إن قرام، إلا أن هذه الطرق في الاستخراج الآلي لا تزال قليلة الفائدة عندما يتعلق الأمر باستكشاف عدد من أنواع التراكيب الاصطلاحية غير المتصلة، أو التي تطرأ عليها تغيرات صرفية ونحوية متنوعة بحسب السياق الذي تكون فيه؛ لأن مثل هذا النوع من التراكيب يتطلب معالجة حاسوبية دقيقة في عدد من مراحل التحليل اللغوي حتى يتمكن النظام الآلي من استخراجه في سياقاته وحالاته المتعددة.

أما بالنسبة لتقنيات الاستخراج والتعرف الآلي المعتمدة على تعلم الآلة، ففي الغالب أنها بحاجة في مرحلة تدريب خوارزميات التصنيف إلى الاعتماد على مصادر لغوية حاسوبية مزودة بمعلومات لغوية في مستويات متعددة، وبناء هذه المصادر في الغالب يتطلب جهوداً بشرية مضنية ويستغرق أوقاتاً طويلة، لذا فإن من أهم المعوقات لهذه الأبحاث تعذر الوصول في أغلب الحالات إلى معاجم شاملة وكافية تمثل التراكيب الاصطلاحية بكل مظاهرها وخصائصها المختلفة، وكل هذا يؤثر سلبياً بشكل أو بـآخر على جودة المخرجات النهائية لهذه التطبيقات.

ومن أهم المشاكل المعقدة كذلك في هذ المجال، عدم وجود إجماع بين المختصين فيما يتعلق بمنهجية التقييم المعتمدة لمهام المعالجة الحاسوبية للتراكيب الاصطلاحية، ووجود سلبيات لأغلب الطرق المستعملة في تقييم النماذج الحاسوبية المختلفة والتي قد تؤثر سلباً في مدى مصداقيته، فعلى سبيل المثال عند الاعتماد على المصادر اللغوية الآلية في تقييم النتائج، غالباً ما تواجهنا مشكلة ندرة هذه المصادر أو ضعفها وعدم شمولها

للتراكيب المستهدفة في عملية الاستخراج أو التعرف الآلي، وكذلك من ناحية أخرى إذا تم الاعتماد على التقييم غير الآلي، والذي يستعين بالخبراء والمختصين لتصنيف النتائج إلى إيجابية أو سلبية، فإنه في مثل هذ الحالات لا يمكننا التقليل من التأثير السلبي لاستعمال الحدس والميول الشخصية في التقييم؛ ولذا فإن الحاجة ملحة في هذ المجال إلى استحداث منهجية واضحة وشاملة لتقييم التطبيقات المختلفة، بالاستناد إلى معايير علمية وعملية يسهل تطبيقها وتعميم نتائجها.

٥- الخاتمة

مع كثرة الأبحاث وتعدد المناهج والطرق المستعملة في اللسانيات الحاسوبية ومعالجة اللغات البشرية يبقى المجال مفتوحاً والأسئلة البحثية مطروحة لتحقيق الهدف الأسمى لعلوم الذكاء الاصطناعي المتنوعة والذي يتمثل في محاولة أنسنة الآلات وتقليل الفجوة بينها وبين البشر من خلال تعزيز طرق التواصل بين الإنسان والآلة، ومحاولة تحسين أداء الآلة أو الحاسوب في أداء المهام المتصلة باستعمال اللغة، وذلك باستثمار ما توفره الآلات من إمكانات وقدرات خارج قدراتنا البشرية المحدودة.

وكما ذكرنا في مقدمة هذا البحث إن التراكيب الاصطلاحية من المشكلات المعقدة بدرجة تعقيد هذه الظاهرة اللغوية في لغتنا، ولا تزال هذه الظاهرة في اللغة العربية بحاجة إلى مزيد من الدراسة والبحث والتحليل، وخاصة في ظل إدراكنا لما تتميز به اللغة العربية من خصائص هندسية بارعة، ومكونات رياضية متميزة، كالجذر والأوزان الصرفية التي قد تساهم في تسريع تقدم الأبحاث في هذا الميدان، وسد الفجوة بين اللغة العربية والمعالجة الحاسوبية للتراكيب الاصطلاحية بمختلف أشكالها وتطبيقاتها.

وعلى الرغم من وجود الكثير من الدراسات المطبقة على اللغة العربية في هذا الميدان، إلا أنها في مجملها لا تقارن بما وصل له البحث في لغات أخرى كالإنجليزية على سبيل المثال؛ ولهذا التأخر أسباب لعل من أهمها قلة الباحثين والمتخصصين في هذا النوع من الأبحاث، وكذلك قلة المؤسسات البحثية التي تُعنى بجمع المتخصصين في اللسانيات وعلوم الحاسوب -وغيرها من التخصصات ذات العلاقة - لبناء فرق بحثية متكاملة، يمكنها الوصول إلى نتائج ذات قيمة معرفية وتقديم إضافات علمية في هذا المجال المهم.

المراجع العربية

ابن عمر، عبد الرزاق (٢٠٠٧) المتلازمات اللفظية في اللغة والقواميس العربية، مجمع الأطرش، تونس.

أبو داود، محمد (٢٠٠٣) معجم التعبير الاصطلاحي في العرية المعاصرة، دار غريب، القاهرة

أبو سعد أحمد (١٩٨٧) معجم التراكيب والعبارات الاصطلاحية العربية القديم منها والمولد، دار العلم للملايين بيروت

إسماعيل، محمود. حسين، مختار الطاهر. الدوش، سيد عوض (١٩٩٦) المعجم السياقي للتعبيرات الاصطلاحية، مكتبة لبنان، بيروت

بشارة، أنطون (٢٠٠٢) معجم التعابير، مكتبة لبنان، بيروت

حافظ، الطاهر عبد السلام هاشم (٢٠٠٤) معجم الحافظ للمتصاحبات العربية، مكتبة لبنان، بيروت.

حجازي، محمود فهمي (١٩٨٠) الجانب السياقي في المعاجم والكتب في مجال تعليم اللغة العربية لغير الناطقين بها، الندوة العالمية الأولى لتعليم العربية لغير الناطقين بها الرياض (ج١) ، (ص٢٣٢-٢٥١).

حسان، تمام (١٩٧٣) اللغة العربية معناها ومبناها، الهيئة المصرية العامة للكتاب، القاهرة

حمادة، سلوى (٢٠٠٩)المعالجة الآلية للغة العربية، دار غريب، القاهرة.

الخولي، محمد علي (١٩٩٨) التراكيب الشائعة في اللغة العربية، دار الفلاح، عمان.

داود، محمد محمد (٢٠١٤) المعجم الموسوعي للتعبير الاصطلاحي في اللغة العربية، دارنهضة مصر، القاهرة

غريم، باولا سانتيان (٢٠١٤) تصنيف مجدد ومجدد للمتلازمات اللفظية العربية، في «المعجمية العربية قضايا وآفاق ج.٢ ، كنوز المعرفة، عمان

غزالة، حسن (١٩٩٣) ترجمة المتلازمات اللفظية. ترجمان، المغرب (٢:٢)

فايد، وفاء كامل (٢٠١٤) المعاجم العربية القطاعية بين التراث والمعاصرة معجم التعابير الاصطلاحية نموذجا. في «المعجمية العربية قضايا وآفاق ج١ كنوز المعرفة، عمان

فايد، وفاء كامل (٢٠٠٧) معجم التعابير الاصطلاحية في العربية المعاصرة، أبو الهول، القاهرة

القاسمي، علي (١٩٧٩) التعابير الاصطلاحية والسياقية ومعجم عربي لها، اللسان العربي، الرباط (مج١٧ ج١ ص١٧-٣٤).

هليل، محمد حلمي (١٩٩٦) الأسس النظرية لوضع معجم للمتلازمات اللفظية العربية، المعجمية العربية (١٢-١٣) تونس.

المراجع الأجنبية

Abdou, A. (2011). Arabic Idioms: a corpus-based study. London: Routledge.

Alghamdi, A. (2018). A computational lexicon and representational model for Arabic multiword expressions, PhD thesis, University of Leeds

Alghamdi, A., & Atwell, E. (2016). An empirical study of Arabic formulaic sequence extraction methods. The 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia: LREC.

Alghamdi, A., & Atwell, E. (2017). Towards Comprehensive Computational Representations of Arabic Multiword Expressions. In International Conference on Computational and Corpus-Based Phraseology (pp. 415-431). Springer, London.

Al-Sabbagh, R., Girju, R., & Diesner, J. (2014). Unsupervised Construction of a Lexicon and a Repository of Variation Patterns for Arabic Modal Multiword Expressions. Proceedings of the 10th Workshop on Multiword Expressions (MWE), 114–123.

Attia, M. and Tounsi, L. (2010). Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon

of Modern Standard Arabic In: Dublin: Technical report, The NCLT Seminar Series, DCU.

Attia, Mohammed A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In Proceedings of FinTAL 2006, pages 87–98, Turku

Atwell, E.S. (1988). Grammatical analysis of English by statistical pattern recognition In: Pattern Recognition. Springer, pp. 626–635.

Baldwin, T. and Kim, S.N. (2010). Multiword expressions. Handbook of Natural Language Processing, second edition. Morgan and Claypool.

Bar, K., Diab, M. and Hawwari, A. (2014). Arabic Multiword Expressions In: Language, Culture, Computation. Computational Linguistics and Linguistics. Springer, pp. 64–81.

Bartsch, S. (2004). Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence. Gunter Narr Verlag.

Berry-Rogghe, G. (1973). The computation of collocations and their relevance in lexical studies. The computer and literary studies.,pp.103–112.

Biber Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999). Longman Grammar of Spoken and Written English. Harlow: Longman.

Carpuat, Marine and Mona Diab. (2010). Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In Proceedings of NAACL/HLT 2010, pages 242–245, Los Angeles, CA.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. and Todirascu, A. (2017). Multiword expression processing: a survey. Computational Linguistics.,pp.1–92.

Cowie, A.P. (1998). Phraseology: Theory, analysis, and applications. OUP Oxford.

da Silva, Joaquim Ferreira, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. (1999). Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In Proceedings of the 9th Portuguese Conference on Artificial Intelligence, pages 113–132, London.

Dale, R. (2010). Classical approaches to natural language processing. In Indurkhya, N. and Damerau, F.J. eds., 2010. Handbook of natural language processing (Vol. 2). CRC Press.

Erman, B. and Warren, B., (2000). The idiom principle and the open choice principle. Text-Interdisciplinary Journal for the Study of Discourse, 20(1), pp.29-62.

Evert, Stefan. (2005). The Statistics of Word Co-occurrences: Word Pairs and Collocations. Ph.D. thesis, University of Stuttgart, Stuttgart.

Farahmand, Meghdad and James Henderson. (2016). Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In Proceedings of the ACL 2016 Workshop on MWEs, pages 61–66, Berlin

Fillmore, C.J. (1979). On fluency In: Individual differences in language ability and language behavior. Elsevier, pp. 85–101.

Francopoulo, G. (2013). LMF lexical markup framework. Hoboken, NJ; London: ISTE Ltd.

Ghoneim, Mahmoud and Mona Diab. (2013). Multiword expressions in the context of statistical machine translation. In Proceedings of IJCNLP 2013, pages 1181–1187, Nagoya

Girju, Roxana, Dan Moldovan, Marta Tatu,and Daniel Antohe. (2005). On the semantics of noun compounds. CSL Special Issue on MWEs, 19(4):479–496.

Green, Spence, Marie-Catherine de Marneffe, and Christopher D. Manning. (2013). Parsing models for identifying multiword expressions. Computational Linguistics, 39(1):195–227.

Haddar, K., & Benhamadou, A. (2010). A Syntactic Lexicon for Arabic Verbs. Information Retrieval, (July 2014), 269–272.

Hashimoto, Chikara and Daisuke Kawahara. (2008). Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In Proceedings of EMNLP 2008, pages 992–1001, Waikiki, HI

Hawwari, A., Attia, M., & Diab, M. (2014). A framework for the classification and annotation of multiword expressions in dialectal arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) (pp. 48-56).

Katz, Graham and Eugenie Giesbrecht. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In Proceedings of the ACL/COLING 2006 Workshop on MWEs, pages 12–19, Sydney

Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. (2016). Parseme survey on MWE resources. In Proceedings of LREC, pp 2299–2306 Portoroz.

Ludeling, A., & Kyto, M. (2008). Corpus linguistics: An international handbook. Walter de Gruyter. Berlin

McCarthy, Diana, Bill Keller, and John Carroll. (2003). Detecting a continuum of compositionality in phrasal verbs. In Proceedings of the ACL 2003 Workshop on MWEs, pages 73–80, Sapporo.

Meghawry, S., Elkorany, A., Salah, A., & Elghazaly, T. (2015). Semantic Extraction of Arabic Multiword Expressions. Computer Science & Information Technology (CS & IT).

Mel'čuk, I., (1998). Collocations and lexical functions. Phraseology. Theory, analysis, and applications, pp.23-53.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990). Introduction to wordnet: An on-line lexical database. International journal of lexicography. 3(4),pp.235–244.

Moirón, M.B.V. (2005). Data-driven identification of fixed expressions and their modifiability. PhD thesis, University of Groningen

Monti, Johanna, Federico Sangati, and Mihael Arcan. (2015). TED-MWE: A bilingual parallel corpus with MWE annotation. In Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015, pages 193–197, Trento.

Najar, D., Mesfar, S. and Ghezela, H. Ben (2015). A large terminological dictionary of Arabic compound words In: International NooJ Conference. Springer, pp. 16–28.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. Formulaic language. 2,pp.387–404.

Pawley, A., &F. Syder. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. J. Richards&R. Schmidt (eds.). Language and Communication, pp.191-226.

Pecina, Pavel. (2008). Lexical Association Measures: Collocation Extraction. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague.

Ramisch, C. (2015). Multiword Expressions Acquisition: A Generic and Open Framework. Springer. London.

Ramisch, Carlos, Aline Villavicencio, Leonardo Moura, and Marco Idiart. (2008). Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In Proceedings of CoNLL 2008, pages 49–56, Manchester

Reddy, Siva, Diana McCarthy, and Suresh Manandhar. (2011). An empirical study on compositionality in compound nouns. In Proceedings of IJCNLP 2011, pages 210–218, Chiang Mai.

Riedl, Martin and Chris Biemann. (2015). A single word is not enough: Ranking multiword expressions using distributional semantics. In Proceedings of EMNLP 2015, pages 2430–2440, Lisbon.

Rikters, M. and Bojar, O. (2017). Paying Attention to Multi-Word Expressions in Neural Machine Translation. Machine Translation Summit XVI, Nagoya, Japan

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP In: Computational Linguistics and Intelligent Text Processing. Springer, pp. 1–15.

Salehi, Bahar, Paul Cook, and Timothy Baldwin. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In Proceedings of NAACL/HLT 2015, pages 977–983, Denver, CO.

Seretan, Violeta. (2011). Syntax-Based Collocation Extraction, Text, Speech and Language Technology, Springer

Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach. John Wiley & Sons.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

Stevens, M.E. and Giuliano, V.E. (1965). Statistical Association Methods for Mechanized Documentation: Symposium Proceedings, Washington, 1964. US Government Printing Office.

Stevenson, Suzanne, Afsaneh Fazly, and Ryan North. (2004). Statistical measures of the semi productivity of light verb constructions. In Proceedings of the ACL 2004 Workshop on MWEs, pages 1–8, Barcelona.

Tan, Liling and Santanu Pal. (2014). Manawi: Using multi-word expressions and named entities to improve machine translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 201–206, Baltimore, MD

Vintar, Š., Vintar, Š., Fišer, D. and Fišer, D. (2008). Harvesting Multi-Word Expressions from Parallel Corpora. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). (Fišer),pp.1091–1096.

Wray, A. 2002. Formulaic language and the lexicon. Cambridge University Press Cambridge.

Zarrieß, S. and Kuhn, J. (2009). Exploiting translational correspondences for pattern-independent MWE identification In: Proceedings of the Workshop on Multiword Expressions Identification, Interpretation, Disambiguation and Applications - MWE '09. Morristown, NJ, USA: Association for Computational Linguistics, pp. 23–30.

Zaghouani, W. (2014). Critical Survey of the Freely Available Arabic Corpora. In Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme In Proceedings of LREC.

فهرس الكتاب

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

مباحث لغوية ١١

خوارزميات الذكاء الاصطناعي في تحليل النص العربي

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المركز حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجتهد في استكتاب نخبة من المحررين والمؤلفين للنهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المركز من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسلة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجهد محرر الكتاب، على ما تفضلوا به من رؤى وأفكار لخدمة العربية في هذا السياق البحثي.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين للتواصل مع المركز؛ لبناء المشروعات العلمية، وتكثيف الجهود، والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

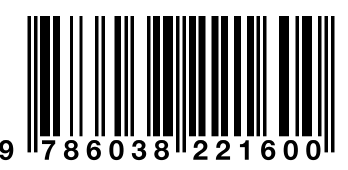
الأمين العام للمركز

أ. د. محمود إسماعيل صالح



ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa




9 786038 221600